Bulk Savings for Bulk Transfers: Minimizing the Energy-Cost for Geo-Distributed Data Centers

Xingjian Lu, Fanxin Kong, Xue Liu, *Member, IEEE*, Jianwei Yin, Qiao Xiang, and Huiqun Yu, *Senior Member, IEEE*

Abstract—With the fast proliferation of cloud computing, major cloud service providers, e.g., Amazon, Google, Facebook, etc., have been deploying more and more geographically distributed data centers to provide customers with better reliability and quality of services. A basic demand in such a geo-distributed data center system is to transfer bulk volumes of data from one data center to another. Geographic distribution and large delay-tolerance of such inter-data-center bulk data transfers provide cloud service providers opportunities to optimize the operating cost. Most existing studies on inter-data-center bulk data transfers focus on minimizing the network bandwidth cost. However, the energy-cost of the bulk data transfers, which also accounts for a large proportion of operating cost in the data centers, still remains unexplored. This is an important problem, especially in the multi-electricity-market environment, where the electricity price exhibits both spatial and temporal diversities. In this paper, we systematically study the problem of how to route and schedule inter-data-center bulk data transfers to minimize the energy-cost for geo-distributed data centers. We model this problem as a min-cost multi-commodity flow problem and develop an efficient two-stage optimization method to solve it. Extensive evaluations with real-life inter-data-center network and electricity prices show that our method brings significant energy-cost savings over existing bulk data transfer methods.

Index Terms—Data center, bulk data transfer, energy-cost, geographical load balance

1 Introduction

The fast development of cloud computing promotes the rapid growth of data centers [1]. To support the expanding scale of cloud applications, major cloud service providers (CSPs), such as Amazon, Google, Facebook, etc., have been deploying tens or even hundreds of geographically distributed (geo-distributed) data centers to provide better reliability and quality of services (QoS). A basic demand in such a geo-distributed data center system is to transfer bulk volumes of data from one data center to another, e.g., periodic data backup, software distribution, virtual machines cloning, etc. [2], [3], [4]. The traffic for such usages is referred to as inter-DC Bulk Data Transfer (BDT) in this paper. The inter-DC BDT is usually of high data volume and delay-tolerant, i.e., CSPs have a span from a few hours to days to finish each transfer. To get a sense of data volumes involved, consider current survey [5] which shows that more 77 percent of data center operators run their regular backup and replication services among three or more data centers, and more than half of CSPs predict that inter-DC BDTs will double or triple over the next couple of years.

Geographic distribution and large delay-tolerance of inter-DC BDTs provide CSPs opportunities to reduce the operating cost. Spatially, data centers at different locations have different operating characteristics, e.g., network bandwidth and unit electricity prices [6]. So CSPs can reduce the operating cost by assigning different routes for different inter-DC BDTs. Temporally, an inter-DC BDT can start at any time after its arrival, as long as it can be completed before the deadline. Thus CSPs can make flexible scheduling for inter-DC BDTs to reduce the operating cost.

Quite some work has been carried out to leverage the spatial and temporal flexibility to optimize the operating cost for inter-DC BDTs, e.g., [2], [3], [4], [7], [8], [9], [10], [11]. These work however, only focuses on optimizing the bandwidth cost, and the energy-cost when performing these inter-DC BDTs has been neglected. The energy-cost takes up as much as the bandwidth cost in a data center, which accounts for around 15 percent [12]. The inter-DC BDT takes up to 45 percent of the total data transfers[7], which has been accounting for 20 percent or more (more than 20 billion US dollars per year) of the energy consumed by the data center [13], [14], [15]. So the inter-DC BDT represents a large portion of energy-cost, and a systematic study on minimizing the energy-cost for inter-DC BDTs has become an important demand for major cloud service providers.

Though dynamic speed scaling [16], energy efficient routing protocols [17], energy-aware data transfer algorithms [18], and Geographical Load Balancing (GLB) techniques

- *X. Lu and H. Yu are with the School of Information Science & Engineering, East China University of Science and Technology, Shanghai 200237, China. E-mail: {luxj, yhq}@ecust.edu.cn.*
- *F. Kong is with the Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104. E-mail: fanxink@seas.upenn.edu.*
- *X. Liu is with the School of Computer Science, McGill University, Montreal, QC H3A 2A7, Canada. E-mail: xueliu@cs.mcgill.ca.*
- *J. Yin is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. E-mail: zjuyjw@zju.edu.cn.*
- *Q. Xiang is with the Department of Computer Science, Yale University, New Haven, CT 06520-8285. E-mail: qiao.xiang@cs.yale.edu.*

Manuscript received 20 June 2016; revised 26 Feb. 2017; accepted 4 Aug. 2017. Date of publication 14 Aug. 2017; date of current version 11 Mar. 2020. (Corresponding author: Xingjian Lu.)
Recommended for acceptance by D. Kliazovich.
Digital Object Identifier no. 10.1109/TCC.2017.2739160

[19], [20], [21], [22] have been developed to reduce the energy-cost of geo-distributed data centers, none of them focuses on minimizing the energy-cost of inter-DC BDTs in the whole inter-DC network. Therefore, we study the novel problem of minimizing the energy-cost of inter-DC BDTs for the geo-distributed data centers in this paper. This problem is equally important and non-trivial because we need to fully explore the spatial and temporal flexibility brought by 1) the operating characteristics (e.g., link capacity) of geographically distributed data centers, 2) the large delay-tolerance of inter-DC BDTs and 3) the time-varying and regional electricity price in the multi-electricity-market environment.

The main contribution of this paper is three fold:

- We study the novel problem of minimizing the energy-cost of inter-DC BDTs for geo-distributed data centers under the multi-electricity-market environment. We formulate this optimization problem (MIN-EC-BDT) in a min-cost multi-commodity flow model.
- We develop a two-stage method to quickly solve the optimal solution to the MIN-EC-BDT. For each BDT, our method searches for the optimal demand division along available time slots in the first stage. It then computes the optimal route and schedule for each portion in the optimal demand division respectively.
- Extensive evaluations with real-life inter-DC network and electricity prices show that our two-stage optimization method can bring significant energy-cost savings over existing inter-DC BDT methods. The results also demonstrate high computation efficiency of our method in saving energy-cost of inter-DC BDTs for geo-distributed data centers.

The rest of this paper is organized as follows. We discuss the related work on inter-DC BDTs in Section 2. We present our system model and formulate the energy-cost minimization problem of inter-DC BDTs for geo-distributed data centers in Section 3. We propose our two-stage optimization method for inter-DC BDTs in Section 4 and evaluate its performance on a real-life inter-DC network with real-life electricity prices in Section 5. We make concluding remarks about our work in Section 6.

2 Related Work

With the development of geo-distributed data centers, the scheduling and routing of inter-DC BDTs have drawn increasing interests from both academia and industry. To optimize data transfers of inter-DC BDTs, Chen et al. [7] conduct the first measurement study on inter-DC traffic characteristics using datasets collected from five major Yahoo! data centers. Leveraging the large delay-tolerance of BDTs and diversity of geo-distributed data centers, researchers have studied the optimal flow assignment for inter-DC BDTs to minimize the bandwidth cost of data centers.

Laoutaris et al. [23] propose transmitting delay-tolerant bulk data for emerging scientific and industrial applications by conservatively utilizing already-paid-for off-peak bandwidth resulting from diurnal data traffic and percentile pricing. They also present the design, implementation, and validation of NetStitcher [3], a system for stitching together unutilized bandwidth across different data centers. Wang et al. [11] formulate the inter-DC BDTs problem into a linear programming model and then iteratively compute the optimal multi-path routing and bandwidth allocation under max-min fairness. Nandagopal et al. [8] propose GRESE. It is an algorithm that leverages the flexibility of large deadlines of inter-DC BDTs to reduce the billable bandwidth usage. Feng et al. [4] present Jetway, a set of algorithms designed to minimize CSP' bandwidth cost on inter-DC video traffic, which has more stringent delays than inter-DC BDTs.

Studies above demonstrate the necessity and importance on understanding characteristics of inter-DC BDTs. However, they mainly focus on optimizing the bandwidth cost. None of them tries to minimize the energy-cost for inter-DC BDTs, which also takes up a large portion of the operating cost in a data center. Therefore, a systematic study on minimizing the energy-cost of inter-DC BDTs for geo-distributed data centers is an urgent task.

A survey of techniques and architectures for designing energy-efficient data centers is introduced in [24], [25]. There is quite some work on energy-efficiency techniques in data centers, such as the dynamic voltage/frequency scaling, energy efficient routing protocols, resource consolidation (virtualization and workload consolidation), and renewable energy resources. Dynamic speed scaling [26] and energy efficient routing methods [17] are developed to save the power for data center devices. Shuja et al. [27] propose a data center-wide energy-efficient resource scheduling framework by scheduling the resources according to current workloads of the data center. Lin et al. [28] use resource consolidation, which can dynamically control the number of activated servers, to save the power in the data center. Shuja et al. [29] and Kong et al. [30] summary studies that focus on using renewable energy and waste heat utilization techniques to improve energy-efficiency for sustainable and green cloud data centers. However, these methods are generally proposed to reduce the energy consumption of a single data center, without considering how to cut the energy-cost for geo-distributed data centers, especially in the multi-electricity-market.

To reduce the energy-cost for geo-distributed data centers, a new scheduling technique called Geographical Load Balancing (GLB) [19], [20], [21], [22] has been developed in recent years. Qureshi et al. [19] are the first to propose the idea of GLB (exploiting geographical and temporal differences in electricity prices). Rao et al. [20] first study the GLB problem as a constrained mixed-integer programming problem. Yao et al. [21] propose a two time scale control algorithm for GLB to achieve energy-cost reductions for the delay-tolerant workloads. Our previous work [22] proposes a joint job scheduling policy and an ADMM-based algorithm to solve the heterogeneity problem of underlying platform and workload demands for GLB. The purposes of [19], [20], [21], [22] are most similar to ours. However, these works mainly focus on the optimal solution for job requests by distributing them to different data centers. It's a load balancing like technique. This paper aims to minimize the energy-cost for inter-DC BDTs, which is a kind of backend traffic between geo-distributed data centers. Moreover, scheduling for jobs just needs to distribute requests among data centers (finding out a single data center to deal with the request), while scheduling for inter-DC BDTs needs to find out one or more paths (may through multiple intermediate data centers) to transfer data from the source data

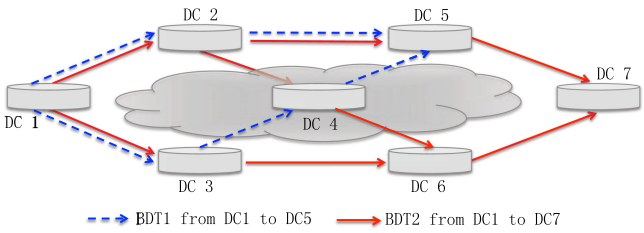


Fig. 1. Inter-DC network and two inter-DC BDTs.

center to the destination data center. It's more complicated to schedule inter-DC BDTs than job requests, since inter-DC BDTs have more spatial and temporal flexibility to be utilized. So the target workload type and optimization objective are different between our method and existing GLB methods.

The inter-DC BDT is also considered troublesome in recent work on distributed data analytics [31], [32]. Mohan et al. [31] present a new system GUPT, which makes privacy-preserving easy for distributed data analysis. Vulimiri et al. [32] propose a solution to optimize the bandwidth cost for wide area analytics over geo-distributed data structured as SQL tables. The optimal scheduling and routing of inter-DC BDTs can be used to guide optimizing query execution plans for distributed data analysis over geo-distributed data centers. However, different from [31], [32], we mainly focus on optimizing data transfers for inter-DC BDTs, instead of optimizing SQL analytics or preserving privacy for geo-distributed data. The target workload type and proposed models and algorithms of this paper are also different from previous works. To the best of our knowledge, we are the first to explicitly address the energy-cost minimization problem of inter-DC BDTs for geo-distributed data centers in the multi-electricity-market environment.

3 System Model and Problem Formulation

In this section, we first present the modeling of inter-DC bulk data transfers, followed by its energy-cost modeling. Then we formulate the energy-cost minimization problem of inter-DC BDTs as a linear programming problem.

3.1 Inter-DC Bulk Data Transfer

We model an inter-DC network as a directed graph $G = (V, E)$. As Fig. 1 shows, each node $i \in V$ represents a data center and each link $(i, j) \in E$ represents the data transfer link from i to j. The whole network operates as a discrete-time system, in which time is divided into a sequence of slots with the same length, denoted by $t \in 1, 2, 3, \ldots$. We use $c_{ij}(t)$ to denote the rate capacity of link $(i, j) \in E$ at time slot t. We define $f_{ij}(t)$ as the actual data flow rate on link (i, j) during time slot t. We use $\mathcal{K}(t)$ to represent the set of BDT requests in the network at the beginning of time slot t. Each bulk data transfer request $k \in \mathcal{K}(t)$ is then defined by a four-tuple (s_k, d_k, dem_k, t_k), where s_k and d_k denote the source and destination data center respectively, dem_k denotes the volume of bulk data, and t_k denotes the deadline by which all dem_k data needs to be transferred from s_k to d_k.

During the transfer process, the data of each BDT request k can be split into multiple parts and transferred to the destination at different time slots. Compared to the length of one time slot, the data transmission time of each link is negligible. Therefore for any request k, if a part of its data leaves the source at a certain time slot, it will be delivered to the destination along one or multiple paths within the same time slot. For example, the BDT1 in Fig. 1 can be transferred from DC1 to DC5 along the paths of DC1-DC2-DC5 and DC1-DC3-DC4-DC5. We use $r_k(t)$ to denote the data flow of request k that is transferred from s_k to d_k during time slot t. We use $f_{ij}^k(t)$ to denote the data flow of BDT request k on link (i, j) at time slot t, and can have $\sum_{k \in \mathcal{K}} f_{ij}^k(t) = f_{ij}(t)$.

3.2 Energy-Cost Modeling for Inter-DC BDTs

Given an inter-DC BDT request, the energy-cost to finish it varies depending on how and when the data is transferred to the destination. The geo-distributed data centers are typically inter-connected with high-capacity links leased from the Internet Service Providers (ISPs). So in terms of energy-cost, CSPs mainly focus on the end systems (source, intermediate, and destination data centers), not the network infrastructure. Thus in this paper, we focus on minimizing the energy-cost of data centers when they performing the inter-DC BDTs. Specifically, we model the total energy-cost of geo-distributed data centers by accumulating the energy-cost of each data center, not by each BDT request. Though the energy-cost of a data center is affected by many factors, e.g., traffic load, temperature, QoS policies and floor space, we model it from the data traffic point of view in this paper. Given a data center i, its BDT traffic flow at time slot t is categorized into incoming data flow and outgoing data flow, denoted as $f_i^{in}(t)$ and $f_i^{out}(t)$, respectively. Using our definition of traffic flow in Section 3.1, the incoming data flow and outgoing data flow can be further expressed as in Eqs. (1) and (2)

$$f_i^{in}(t) = \sum_{k \in \mathcal{K}} \sum_{j \in V} f_{ji}^k(t), \tag{1}$$

$$f_i^{out}(t) = \sum_{k \in \mathcal{K}} \sum_{j \in V} f_{ij}^k(t). \tag{2}$$

We then use $e_i(t)$ to denote the energy consumption of data center i during time slot t. It is expressed as the sum of energy consumption incurred by the incoming and outgoing bulk data flow at data center i, as shown in

$$e_i(t) = e_i^{in}(f_i^{in}(t)) + e_i^{out}(f_i^{out}(t)), \tag{3}$$

where $e_i^{in}(\xi)$ and $e_i^{out}(\xi)$ are energy consumption functions for receiving and sending inter-DC BDT flow of value ξ at data center i, respectively. We assume both $e_i^{in}(\xi)$ and $e_i^{out}(\xi)$ are proportional to ξ, i.e., $e_i^{in}(f_i^{in}(t)) = e_i^{in} f_i^{in}(t)$ and $e_i^{out}(f_i^{out}(t)) = e_i^{out} f_i^{out}(t)$. Here e_i^{in} and e_i^{out} represent the energy consumption per incoming and outgoing BDT flow at data center i, respectively. Although the workload proportional energy consumption is relatively hard to achieve on a standalone server because of hardware constraints, it is possible to achieve energy proportionality on a data center as we can control the number of active and inactive nodes [33]. Fine-grain resource management and server consolidation using virtual machines or container systems further

allow for energy proportionality in data centers [34]. So it's a reasonable simplifying assumption for workload proportional energy consumption of data centers.

The study of more precise definition of functions $e_i^{in}(\xi)$ and $e_i^{out}(\xi)$ is out of this paper's scope and will be explored as our future work.

We then use $a_i(t)$ to denote the electricity price that data center i needs to pay for data transfer flow in time slot t. With the energy consumption function and electricity price defined, we are able to express the monetary energy-cost incurred by inter-DC BDT flows at data center i at time slot t as

$$\varphi_i(t) = e_i(t)a_i(t). \tag{4}$$

3.3 Problem Formulation

In this paper, we aim to find the optimal routing and scheduling for inter-DC BDTs to minimize the energy-cost of geo-distributed data centers in the multi-electricity-market with time-varying and regional electricity prices. Towards this objective, we model this problem as a min-cost multi-commodity flow problem in a dynamic network.

Given an inter-DC network $G = (V, E)$ operated in finite equal-time slots, i.e., $t \in \mathbb{T} = \{1, 2, \ldots, T\}$, a time-varying capacity $c_{ij}(t)$, is assigned to each link (i, j). A time-varying electricity price $a_i(t)$ is assigned to each node i. At the beginning of time slot t_0, a set of inter-DC BDT requests $\mathcal{K}(t_0)$ is received and scheduled. Without loss of generality, we assume all the inter-DC BDT requests in $\mathcal{K}(t_0)$ have the same deadline, i.e., $t_k = T$, for any request $k \in \mathcal{K}(t_0)$.

Our decision variables are $f_{ij}^k(t)$, the actual data flow of each BDT request k along each link (i, j) at every time slot t, and $r_k(t)$, the data volume of each BDT request k transferred from s_k to d_k at every time slot t. Given this dynamic inter-DC network and all the decision variables, our objective is to route and schedule all the BDT requests $\mathcal{K}(t_0)$ to be finished before the end of time slot T such that the energy-cost of all the geo-distributed data centers to accomplish these BDT requests is minimized.

Combining all the variables, constraints and the objective function, we formally define the following minimizing energy-cost of bulk data transfers (MIN-EC-BDT) problem

MIN-EC-BDT:

$$\min_{f_{ij}^k(t), r_k(t)} \sum_{t \in \mathbb{T}} \sum_{i \in V} \varphi_i(t), \tag{5}$$

s.t.

$$\varphi_i(t) = \left(e_i^{in} \sum_{k \in \mathcal{K}(t_0)} \sum_{j \in V} f_{ji}^k(t) + e_i^{out} \sum_{k \in \mathcal{K}(t_0)} \sum_{j \in V} f_{ij}^k(t) \right) a_i(t), \tag{6}$$

$$\sum_{k \in \mathcal{K}(t_0)} f_{ij}^k(t) \le c_{ij}(t), \forall t \in \mathbb{T}, \forall (i, j) \in E, \tag{7}$$

$$\sum_{j \in V} f_{ij}^k(t) - \sum_{j \in V} f_{ji}^k(t) = \begin{cases} r_k(t) & \text{if } i = s_k \\ 0 & \text{if } i \in V / \{s_k, d_k\} \\ -r_k(t) & \text{if } i = d_k \end{cases} \tag{8}$$

$$\forall k \in \mathcal{K}(t_0), \forall t \in \mathbb{T},$$

$$\sum_{t \in \mathbb{T}} r_k(t) = dem_k, \quad \forall k \in \mathcal{K}(t_0), \tag{9}$$

$$r_k(t) \ge 0, \forall k \in \mathcal{K}(t_0), \forall t \in \mathbb{T}, \tag{10}$$

$$f_{ij}^k(t) \ge 0, \forall (i, j) \in E, \forall k \in \mathcal{K}(t_0), \forall t \in \mathbb{T}. \tag{11}$$

Constraint (6) represents the energy-cost of data center i at time slot t, which is derived from Eqs. (1), (2), (3), and (4). Constraint (7) is the link capacity constraint, which means the total flow along a link cannot exceed its capacity at any time slot. Constraint (8) represents the flow conservation constraint. It ensures that for any request k, the partial data $r_k(t)$ leaves the source s_k at time slot t will be delivered to d_k within the same time slot. For each inter-DC BDT request, constraint (9) ensures that it is fulfilled by the end of time slot T. Constraints (10) and (11) are non-negative constraints of partial data transfer per request and actual data flow per link, respectively.

3.4 Model Discussions

We now discuss the assumptions made in above models and some practical considerations.

Above, we assume the system operates in slotted time, which has been widely used as a reasonable solution to convert the complex continuous time model into discrete optimal decision process. The bulk data flows across inter-DC links can be split and transmitted along multiple multi-hop paths, each of which can be optimally computed over time. It is a common assumption and made by [3], [4], [9], [11].

The time-varying regional electricity prices are assumed to be known in advance when making the scheduling decision at each time interval. Such an assumption is commonly made in the literature [35], [36]. Some statistical machine learning techniques [37], [38] can be used to achieve such kind of predictions. Furthermore, due to huge electricity demands, the data center acquires electricity from girds using long term contracts in day ahead market, since the long term contracts cost lower than the real time market price of electricity [39]. Such day-ahead hourly or 15-minute price information can be also used to as or guide the electricity price prediction.

We study the power of a data center only from a data traffic load point of view, without considering other factors, e.g., temperature. Because the traffic load is the most important (and we can control) operating characteristic that affects the power of inter-DC BDTs. Though we focus on inter-DC BDT power modelling from the data traffic load point of view, important to note that such consideration has no impact on the essence of the proposed algorithm.

Finally, in this work, for highlighting the key points of our method, we just consider the energy-cost when making the optimal routing and scheduling decisions for inter-DC BDTs. However, our model and algorithms can be easily extended to accommodate other kinds of operating cost. It shows that our method can be also applied to achieve other kind of optimizations for inter-DC BDTs.

4 Solution Algorithm Design

The MIN-EC-BDT is an typical optimization and scheduling problem for dynamic network flow. In this section, we first describe and analyze the time-expansion based approach

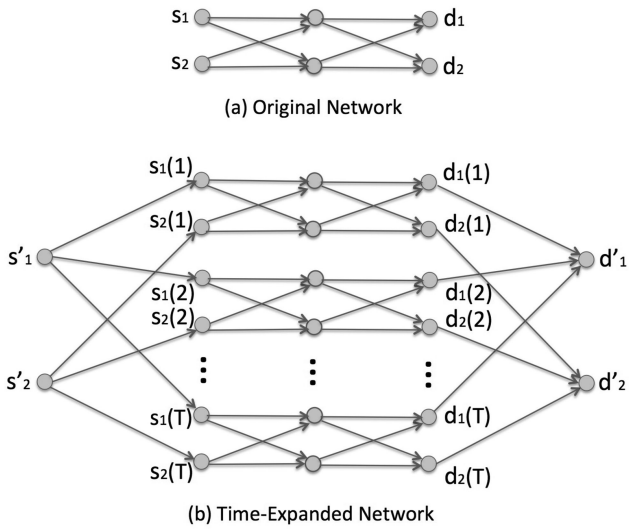


Fig. 2. Transformation between original network and time-expanded network.

for solving the MIN-EC-BDT problem. Then a two-stage optimization method for this problem is proposed.

4.1 Time-Expansion Based Approach

One of the most straightforward methods to solve flow problems on a dynamic network is to reduce them to similar problems on a static time-expanded network [40], [41]. By introducing a virtual copy of all nodes at each time interval, it can translate this dynamic problem into an equivalent static problem in a time-expanded graph.

Given a dynamic inter-DC network $G = (V, E)$ with negligible link transmission time, the time-expanded network $G^T = (V^T, E^T)$ for inter-DC BDTs can be constructed as follows: First, T copies of N are created, each of which corresponds to an instance of G at time slot t. Then, for each request $k \in \mathcal{K}(t_0)$, a super source node $s_k^{'}$ and a super destination node $d_k^{'}$ are added. These super source and super destination nodes connect to corresponding source and destination nodes at each copy of the original (also called underlying) network respectively. The capacity of the θth copy of link $e \in E$ corresponds to time-varying link capacity at time slot $\theta = 1, 2, \ldots, T$. The capacity of the new added links that connect with the super nodes is unlimited, and the electricity price of each super node is assumed to be zero.

Fig. 2 shows the time-expanded network for a dynamic network, which consists of six vertexes and eight edges and needs to transmit two BDTs: (s_1, d_1, dem_1, t_1) and (s_2, d_2, dem_2, t_2), where $t_1 = t_2 = T$. With the time-expanded network, the MIN-EC-BDT can be tackled by solving a typical min-cost multi-commodity flow problem on the static network.

The time-expansion based approach simplifies the considered flow problem and makes available algorithmic toolbox developed for static flows. However, the size of a time-expanded network is typically very large for realistic problems. For example, assuming an inter-DC network with 100 nodes and the length of a time slot is 5 minutes, a typical deadline $T = 24$ hours involves a matrix with around $8.3 * 10^8$ elements. As we will show in Section 5.3.2, the performance of the time-expansion based approach is unacceptable, especially when the deadline T is large.

4.2 Two-Stage Optimization Approach

From the formulation of the MIN-EC-BDT (Eqs. (5), (6), (7), (8), (9), (10), and (11)), we observe that if the partial data demand $r_k(t)$ of each request k at each time slot t is determined, the MIN-EC-BDT can be decomposed to T independent min-cost multi-commodity flow problems, which can be efficiently solved on the underlying network in a parallel fashion. It is even more intuitive to see this observation in the time-expanded network in Fig. 2: if all the flows on the links connecting super and original nodes are determined, the time-expanded network can be reduced to T independent copies of the underlying network.

Inspired by the observation above, we consider an efficient two-stage approach to solve the MIN-EC-BDT on the underlying network. Instead of solving the problem by choosing $r_k(t)$ and $f_{ij}^k(t)$ simultaneously, we choose them sequentially. In the first stage, we solve the optimal demand division $r = \{r_k(t)\}$. Then, in the second stage, we compute the optimal flow $f_{ij}^k(t)$ by solving the network flow problem (BDT-Opt-Underlying, described later) on the underlying network with given optimal demand division.

Before stepping into details of the proposed approach, we first illustrate how to derive an upper bound for the flow of each BDT request at each time slot t. To be specific, we need to find the maximum demand satisfaction rate z for each BDT request, such that up to z rate of their demands can be assigned on links at that time slot t. To fully utilize the network resources, we use the optimal max-min fair multi-transfer (OPT-MMF-MT) algorithm to strike the balance between fairness and network utilization, instead of using a common demand satisfaction rate z for all the BDTs. The Max-Min Fair Linear Programming (MMF-LP) model is defined as follows:

MMF-LP:

$$\max_{f_{ij}^k(t)} z, \tag{12}$$

s.t.

$$\sum_{k \in \mathcal{K}(t_0)} f_{ij}^k(t) \le C_{ij}(t), \forall (i,j) \in E, \tag{13}$$

$$\sum_{j \in V} f_{ij}^k(t) - \sum_{j \in V} f_{ji}^k(t) = \begin{cases} z * dem_k & \text{if } i = s_k \\ 0 & \text{if } i \in V/\{s_k, d_k\} \\ -z * dem_k & \text{if } i = d_k \end{cases} \quad \forall k \in \mathcal{K}_{unsat}(t_0), \tag{14}$$

$$\sum_{j \in V} f_{ij}^k(t) - \sum_{j \in V} f_{ji}^k(t) = \begin{cases} z_{sat}^k * dem_k & \text{if } i = s_k \\ 0 & \text{if } i \in V/\{s_k, d_k\} \\ -z_{sat}^k * dem_k & \text{if } i = d_k \end{cases} \quad \forall k \in \mathcal{K}_{sat}(t_0), \tag{15}$$

$$f_{ij}^k(t) \ge 0, z \ge 0, \forall (i,j) \in E, \forall k \in \mathcal{K}(t_0). \tag{16}$$

The objective of MMF-LP is to maximize the demand satisfaction rate z for unsaturated BDT requests ($k \in \mathcal{K}_{unsat}(t_0)$) and keep the satisfaction rate constant when the requests are saturated ($k \in \mathcal{K}_{sat}(t_0)$). The pseudo code of the OPT-MMF-MT algorithm is shown in Algorithm 1. The basic

idea of this algorithm is to iteratively maximize the satisfaction rate z until all the requests are saturated. In each iteration, it first solves MMF-LP, then finds and removes the saturated requests from subsequent processes by fixing their rate values. The algorithm finally outputs the flow $f_{ij}^k(t)$ and demand satisfaction rate $z_k(t)$ for each request k at every time slot t.

The time complexity of Algorithm 1 is as follows: there two loops in Algorithm 1. For the while loop, at the worst case, line 4 is executed $|k(t_0)|$ times and thus solves MMF-LP $|k(t_0)|$ times. For the for loop, at the worst case, line 7 to line 16 are executed and thus $|k(t_0)|^2$ times and thus solve MMF-LP $|k(t_0)|^2$ times. In sum, MMF-LP is solved $|k(t_0)|^2 + |k(t_0)|$times. Hence, using big O notation, the complexity is that Algorithm 1 solves MMF-LP $O(|k(t_0)|^2)$ times.

Algorithm 1. OPT-MMF-MT Alg. for Max Desired Flow [11]

Input: $\mathcal{K}(t_0)$, $G = (V, E)$.
/* $\mathcal{K}(t_0)$: BDT request set;
G: dynamic inter-DC network at time slot t; */
Output: $f_{ij}^k(t)$, $z_k(t)$.
/* $f_{ij}^k(t)$: Maximum flow on network N at time sot t;
$z_k(t)$: Maximum satisfaction rate for request k at t; */
1. $\mathcal{K}_{sat}(t_0) \leftarrow null$;
2. $\mathcal{K}_{unsat}(t_0) \leftarrow \mathcal{K}(t_0)$;
/* initialization*/
3. **while** $\mathcal{K}_{unsat}(t_0) \neq null$ **do**
4. solve MMF-LP using G, $\mathcal{K}_{sat}(t_0)$, $\mathcal{K}_{unsat}(t_0)$;
5. output $z(t)$, $f_{ij}^k(t)$, $\forall k \in \mathcal{K}(t_0), \forall (i,j) \in E$;
6. **for** $k_i \in \mathcal{K}_{unsat}(t_0)$ **do**
7. **if** k_i has no flow on the residual network of N **then**
8. solve MMF-LP using $\mathcal{K}_{unsat}(t_0) \leftarrow \{k_i\}$, $\forall k \in \mathcal{K}(t_0) \backslash \{k_i\} : z_{sat}^k \leftarrow z(t)$, $\mathcal{K}_{sat}(t_0) \leftarrow \mathcal{K}_{sat}(t_0) \cup \mathcal{K}_{unsat}(t_0) \backslash \{k_i\}$;
9. output $z_{temp}(t)$, $f_{ij}^k(t)$, $\forall k \in \mathcal{K}(t_0), \forall (i,j) \in E$;
10. **if** $z_{temp}(t) = z(t)$ **then**
11. $\mathcal{K}_{sat}(t_0) \leftarrow \mathcal{K}_{sat}(t_0) \cup \{k_i\}$;
12. $\mathcal{K}_{unsat}(t_0) \leftarrow \mathcal{K}_{unsat}(t_0) \backslash \{k_i\}$;
13. $z_{sat}^{k_i}(t) = z(t)$;
14. $z_{k_i}(t) = z_{sat}^{k_i}(t)$;
15. **end if**
16. **end if**
17. **end for**
18. **end while**

Note that the MMF-LP model and the OPT-MMF-MT algorithm focus on an independent time slot t. So they can be implemented in a parallel fashion to find out these upper bounds simultaneously. After deriving all the upper bounds of $r_k(t)$, we need to search for the optimal value for each $r_k(t)$. The optimal demand-division problem is defined as follows:

Demand-Division:

$$\min_{r_k(t)} \Psi_r, \tag{17}$$

s.t.

$$\sum_{t \in \mathbb{T}} r_k(t) \geq dem_k, \forall k \in \mathcal{K}(t_0), \tag{18}$$

$$0 \leq r_k(t) \leq z_k(t) * dem_k, \forall k \in \mathcal{K}(t_0), \forall t \in \mathbb{T}. \tag{19}$$

This problem aims to compute an optimal demand division r, such that the total energy-cost is minimized. Constraint (18) ensures that the data amount of each BDT request should be transferred from the source to the destination before the deadline. Constraint (19) illustrates the value range of each $r_k(t)$. Note that the objective function Ψ_r is complicated. Its value is defined as the objective value of the following energy-cost optimization scheduling problem (BDT-Opt-Underlying) with given demand division r.

BDT-Opt-Underlying:

$$\Psi_r = \min_{f_{ij}^k(t)} \sum_{t \in \mathbb{T}} \sum_{i \in V} \varphi_i(t), \tag{20}$$

s.t.

$$\varphi_i(t) = \left(e_i^{in} \sum_{k \in \mathcal{K}(t_0)} \sum_{j \in V} f_{ji}^k(t) + e_i^{out} \sum_{k \in \mathcal{K}(t_0)} \sum_{j \in V} f_{ij}^k(t) \right) a_i(t), \tag{21}$$

$$\sum_{k \in \mathcal{K}(t_0)} f_{ij}^k(t) \leq c_{ij}(t), \forall (i,j) \in E, \forall t \in \mathbb{T}, \tag{22}$$

$$\sum_{j \in V} f_{ij}^k(t) - \sum_{j \in V} f_{ji}^k(t) = \begin{cases} r_k(t) & \text{if } i = s_k \\ 0 & \text{if } i \in V/\{s_k, d_k\} \\ -r_k(t) & \text{if } i = d_k \end{cases} \quad \forall k \in \mathcal{K}(t_0), \forall t \in \mathbb{T}, \tag{23}$$

$$f_{ij}^k(t) \geq 0, \forall (i,j) \in E, \forall k \in \mathcal{K}(t_0), \forall t \in \mathbb{T}. \tag{24}$$

Though the BDT-Opt-Underlying aims to solve a dynamic flow $f_{ij}^k(t)$, it can be decomposed into an independent min-cost multi-commodity flow subproblem for each time slot t with given partial demand $r_k(t)$. By solving these subproblems separately, the optimal value of Ψ_r can be derived. The dynamic flow $f_{ij}^k(t)$ that produces the optimal value of Ψ_r is the optimal solution of the MIN-EC-BDT.

Our complete two-stage approach for solving the MIN-EC-BDT problem is shown in Algorithm 2. It takes the BDT request set $\mathcal{K}(t_0)$, dynamic inter-DC network G, time-varying electricity prices $a_i(t)$, and the makespan of time slot $\mathbb{T}$ as input, and finally outputs the optimal dynamic flow $f_{ij}^k(t)$ and the minimum cost Ψ_{min}. This algorithm starts with computing the maximum satisfaction rate $z_k(t)$ for each request k at every available time slot t by solving Algorithm 1 (Lines 3-6). Then it iteratively looks for a feasible demand division such that the energy-cost (solved by the BDT-Opt-Underlying) for that division is minimized (Lines 7-15). The minimized energy-cost is the final output variable Ψ_{min}, and the derived optimal flow is the final optimal solution $f_{ij}^k(t)$. Compared to time-expansion based approach, our method is not only faster from computational point of view, but also avoids explicit space and time expansions. The parallel implementation further reduces the solving time as we will demonstrate in Section 5.3.

The time complexity analysis of Algorithm 2 is as follows: Algorithm 2 calls Algorithm 1 in the for loop and executes Algorithm 1 $|T|$ times. And thus, line 3 to line 6 solve MMF-LP $O(|T||k(t_0)|^2)$ times. Line 7 to line 13 solve BDT-Opt-Underlying $|T||k(t_0)|$ times at most. Further, both

Fig. 3. The U.S. electricity market [43] and used inter-DC network topology.

MMF-LP and BDT-Opt-Underlying have $|E|$ variables and both are linear programming problems. Hence, the overall complexity of Algorithm 2 is $O(|T||k(t_0)|^2)T(LP)$, where $T(LP)$ is the complexity for solving linear programing problems. We know that linear problems can be solved in polynomial time and thus Algorithm 2 has a polynomial time complexity.

Algorithm 2. Complete Alg. for the MIN-EC-BDT

Input: $\mathcal{K}(t_0)$, $N=(V,E)$, $a_i(t)$, $\mathbb{T}$.
/* $\mathcal{K}(t_0)$: BDT request set at scheduling period t_0;
G: dynamic inter-DC network;
$a_i(t)$: time-varying electricity price for data center i;
$\mathbb{T}$: makespan of time slots that need to schedule; */
Output: $f_{ij}^k(t)$, Ψ_{min}.
/* $f_{ij}^k(t)$: optimal dynamic flow on network N;
φ_{min}: the minimum total energy-cost; */
1. $\varphi_{min} \leftarrow \infty$;
2. $f_{ij}^k(t) \leftarrow$ zero flow;
3. **for** $t \in \mathbb{T}$ **do**
4. solve Algorithm 1 using $\mathcal{K}(t_0)$, G at time slot t;
5. output $z_k(t)$ for each request k;
6. **end for** /* solving the maximum satisfaction rate for each request k at each time slot t */
7. **while** exist demand division r s.t. Eqs. (18), (19) **do**
8. solve BDT-Opt-Underlying using r, G, $a_i(t)$;;
9. output Ψ_r, $x_{ij}^k(t)$;
10. **if** $\Psi_r \leq \varphi_{min}$ **then**
11. $\Psi_{min} \leftarrow \Psi(r)$;
12. $f_{ij}^k(t) \leftarrow x_{ij}^k(t)$;
13. **end if**
14. remove this division from feasible solution space of Demand-Division problem;
15. **end while**

5 Evaluation

In this section, we evaluate the proposed two-stage optimization method for the MIN-EC-BDT with real-life inter-DC network and electricity prices.

5.1 Evaluation Settings

To characterize the benefits on energy-cost savings brought by the spatial and temporal flexibility of inter-DC BDTs, we perform simulation on a large inter-DC network, which is composed of 11 geo-distributed data centers with a real-life network topology [42] as shown in Fig. 3. All the geo-distributed data centers are operated by a same cloud service provider. The backbone link between any two adjacent data centers is bidirectional, and the bandwidth is different (varies from 1 to 3 Gbps) for each backbone link, depending on the used network service provider. These data centers reside in different regional electricity markets, e.g., California, Midwest and etc. We use the hourly day-ahead electricity price ($/MWh) of these geo-distributed data centers of a period of 48 hours, i.e., from 0am on Jan 31st, 2012 to 0am Feb 2th, 2012 [43], for our simulation. For each data center, the unit incoming and outgoing energy consumption e_i^{in} and e_i^{out} for inter-DC BDTs are uniformly random between [10, 50] KWh. To match the hourly electricity price data, we set the scheduling interval (time slot) is one hour for inter-DC BDTs, i.e., we can transfer 9000 G data on a link with bandwidth 2.5 Gbps per time slot. The simulation parameters used in this paper are summarized in Table 1.

TABLE 1
Summary of Simulation Parameters

Parameter	Value
N	11 geo-distributed data centers, topology as shown in Fig. 3.
c_{ij}	The bandwidth varies from 1 Gbps to 3 Gbps.
$a_i(t)$	The hourly day-ahead electricity price trace from 0am on Jan 31st, 2012 to 0am Feb 2th, 2012 [35].
e_i^{in}	Uniformly random between [10, 50] KWh.
e_i^{out}	Uniformly random between [10, 50] KWh.
t_0	Scheduling start time from 1 to 24.
$\mathcal{K}(t_0)$	3, 5, 10 inter-DC BDT requests.
d	The demand is 4500 G, 13500 G, 27000 G.
T	The largest deadline is 24.

5.2 Comparison Approaches

We comparatively study the following inter-DC BDT scheduling methods:

- FAST, this method aims to achieve the fastest inter-DC BDT transfer by ensuring the maximum amount of data being transferred in each nearest time slot. This strategy is adopted in [3], [11].
- FAST_MIN, this method minimizes the energy-cost while maintains the fastest transfer. The similar model (maximum concurrent flow and min-cost multi-commodity flow) is previously used in [4] to minimize the bandwidth cost for inter-DC video traffic.
- AVG_Demand, this method evenly distributes the total BDT demand across all the available time slots and then looks for the optimal routing and scheduling for each time slot to minimize the energy-cost.
- EXPANSION, this method is discussed in Section 4.1, which solves the MIN-EBC-BDT by transforming it to a time-expanded network and solving the large-scale transformed problem in the time-expanded graph.
- 2Stage_MinEC, the two-stage optimization method we propose in Section 4.2 to quickly solve the MIN-EC-BDT.
- 2State_MinE, a variation of our two-stage optimization method, which minimizes the energy consumption instead of the energy-cost for inter-DC BDTs.

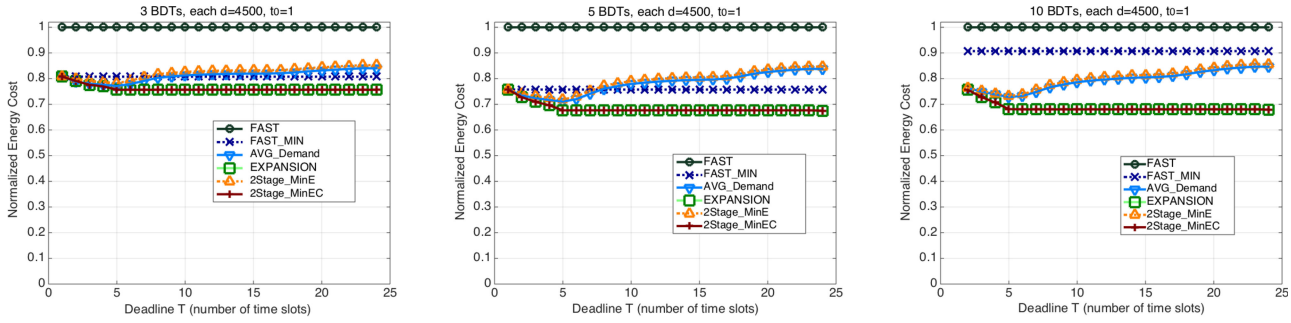


Fig. 4. Normalized energy-cost for all the comparison methods when $d = 10, t_0 = 1$, and the number of BDTs is 3, 5, and 10 separately.

5.3 Evaluation Results

5.3.1 Energy-Cost Savings

First, we set the bulk data demand $d = 4500$ G for each inter-DC BDT, and let the service provider to schedule 3, 5, and 10 inter-DC BDTs respectively.[1] For each number of inter-DC BDTs, the deadline T varies from the minimal value that enables to complete the data demand to the largest deadline 24 used in this evaluation. The scheduling starts at the beginning of time slot $t_0 = 1$, i.e., 0am to 1am on Jan 31st. The normalized energy-cost of different methods is plotted in Fig. 4, by setting the largest energy-cost to 1.

Overall, our method (2Stage_Min_EC) and the EXPANSION method achieve minimum energy-cost in all the experiments. The two methods are equivalent in essence, and the only difference between them is that the 2Stage_Min_EC solves the optimal solution on the original (underlying) network, while the EXPANSION solves the solution on the time-expanded network. It shows a great potential of our method on energy-cost savings for inter-DC BDTs. The energy-cost of our method decreases as T increases until to be converged. It is a desirable property and demonstrates the necessity and importance of leveraging the temporal flexibility to schedule inter-DC BDTs. When T increases, our method has more temporal flexibility to let the inter-DC BDTs being scheduled at time slots with lower electricity prices. So the energy-cost of them decreases. However, when T increases to a certain value, e.g., 5 in Fig. 4a, continually relax it will not bring a lower value, since no lower electricity price exists at the increased time slots.

The FAST method has a constant and largest energy-cost. This is because it always fully utilize nearest time slots to schedule inter-DC BDTs without using any temporal flexibility, resulting in unchanged and much higher energy-cost even T is relaxed. Similarly, the FAST_MIN method, which searches for optimal solution with minimized energy-cost for each nearest time slot, has a constant energy-cost, lower than the FAST method, but still higher than ours. The AVG_Demand and 2Stage_MinE methods have similar and more variable trends on energy-cost. The AVG_Demand method simply utilizes all the temporal flexibility by evenly distributing BDT demands across all the time slots, no matter electricity price is high or low. So its energy-cost depends much on the diurnal variation of electricity prices. Instead of equal division, the 2Stage_MinE method tries to solve an optimal demand division across all the time slots. But it minimizes the energy consumption, not the cost. So its energy-cost is also variable and depends much on electricity prices. Similar to the AVG_Demand method, the energy-cost of the 2Stage_MinE method decreases first ($T \leq 5$) and then increases. The reason is that the electricity prices are often lower in the early morning and then increases.

5.3.2 Computation Time

We also record the computation time (with Matlab 2011 and CVX solver) for each method to get the solution during all the experiments. We plot the average computation time when 5 BDTs and each $d = 13500$ in Fig. 5 as a representative result for comparison.

From this figure we can see that the FAST method and the FAST_MIN method takes the least time to get the solution, not yet considering the parallel implementation of our method (P2Stage_MinEBC). The reason is that the two methods simply fully utilize the nearest time slots to transfer the BDT demand, which needs fewer computation time to find out the solution and the computation time is independent on T (no evident change as T varies). The FAST_MIN method is a little slower than the FAST method, as it still needs some time to find out the solution with minimized energy-cost from the feasible solutions solved by the FAST method. The AVG_Demand and the two two-stage optimization methods (2Stage_MinE and 2Stage_MinEC) have nearly linear computation time with the value of T.

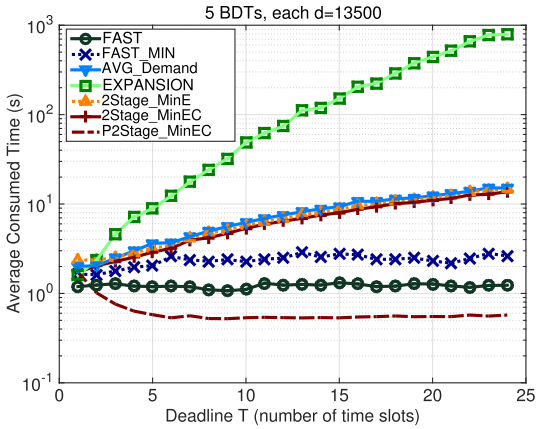


Fig. 5. Average computation time for each method with different deadlines.

1. Since the scale of the used real-life inter-DC network is limited, here the largest number of inter-DC BDTs is sufficiently to set 10. More large-scale evaluation will be explored in Section 5.3.4.

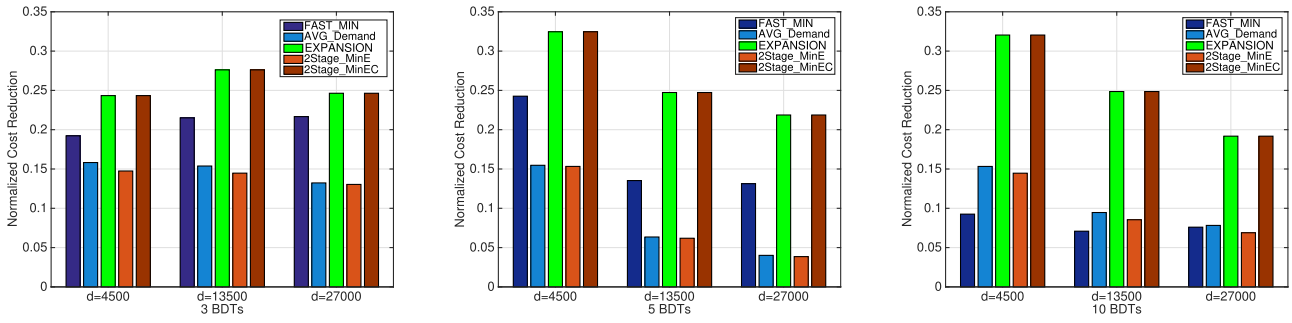


Fig. 6. Energy-cost reduction ratios for all the comparison methods with different number of BDTs when $T = 24$ and $t_0 = 1$.

Though the differences of the three methods on computation time are small, our method (2Stage_MinEBC) takes lower time to find out the optimal solution than the AVG_Demand and 2Stage_MinE methods. Among all the methods, the EXPANSION method takes the largest computation time (nearly exponential to T), since it solves the optimization problem on the time-expanded network, which is much larger than the original network.

Although our method (2Stage_MinEC) is slightly slower than the FAST method, it is faster than other methods and achieves the largest energy-cost savings. The EXPANSION method also achieves the same energy-cost savings as our method, but with a much larger computation time, especially when T is large. So this experiment shows that our method can achieve better balance between energy-cost savings and computation time. Furthermore, the parallel implementation (P2Stage_MinEC) of our method, which uses T parallel threads to find out the optimal solution concurrently, further significantly reduces the computation time (even lower than the FAST method).

5.3.3 Robustness on Data Transfer Volume

In the following, we consider different combinations of the number of inter-DC BDTs 3, 5, and 10, and each inter-DC BDT demand $d = 4500, 13500, 27000$ G, to represent the light, medium and high data transfer volume and evaluate the robustness of our method on the data transfer volume.

By repeating above experiments with different combinations, similar results can be found to $d = 4500$ in Section 5.3.1. The energy-cost of our method is always the lowest. It demonstrates the robustness of our method on data transfer volume. We also plot the energy-cost reduction ratios compared with the FAST method in Fig. 6 when $T = 24$, for each group of d and the number of inter-DC BDTs. As it shows, the energy-cost reduction ratio of our method is significantly higher than that of other methods. The reduction ratio of our method can reach 32.5 percent when the data transfer volume is in the low level (5 BDTs and $d = 4500$ G). Even for high level of the data transfer volume (10 BDTs and $d = 27000$ G), our method can also achieve around 19.2 percent reduction on energy-cost, compared to the FAST method.

Fig. 6 also shows a slight trend that the energy-cost saving achieved by our method decreases as data transfer volume increases, e.g., for a fixed number of inter-DC BDTs (or d), the energy-cost reduction ratio decreases as the value of d (or the number of inter-DC BDTs) increases. The reason is that the link capacity of the inter-DC network is limited, larger data transfer volume generally needs more time slots to transfer the data. Though our method has already given priority to the time slots with lower electricity prices, the energy-cost per BDT data increases as data transfer volume is larger. This is why the energy-cost saving of our method decreases when data transfer volume is larger, even though it has already achieved the minimum energy-cost for each given inter-DC BDT data transfer volume.

5.3.4 Robustness on Scheduling Start Time

In the following, we first fix d and T to 4500 G and 6 respectively, and plot the normalized energy-cost reduction ratio curves in Fig. 7 when t_0 varies from 1 to 24. This figure intuitively shows that different scheduling start time t_0 has a great influence on the energy-cost of inter-DC BDTs. Even for the FAST and FAST_MIN methods, the energy-cost of them are not constant any more. To validate the energy-cost saving robustness of our method on different scheduling start time, we repeat above experiments illustrated in Figs. 4 and 6 by setting t_0 from 1 to 24. For each value of t_0, the energy-cost of our method is still the lowest, which shows that the benefits of our method on energy-cost savings is robust on scheduling start time t_0. The general trends of all the methods are similar to the case of $t_0 = 1$, described before. For comparison with the case of $t_0 = 1$ (free hour and electricity prices are low), we plot the energy-cost and corresponding reduction ratio in Figs. 8 and 9 respectively as $t_0 = 18$ (busy hour and electricity prices are high).

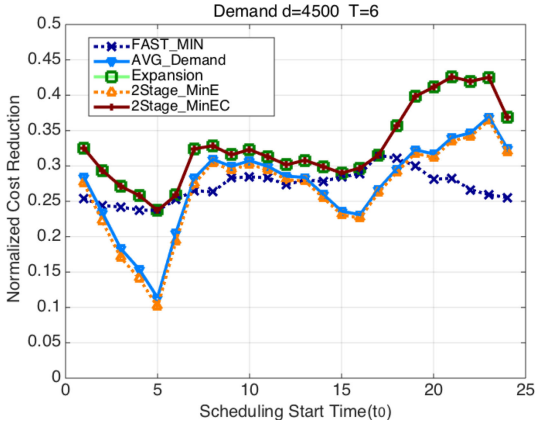


Fig. 7. Normalized energy-cost reduction for different scheduling start time t_0.

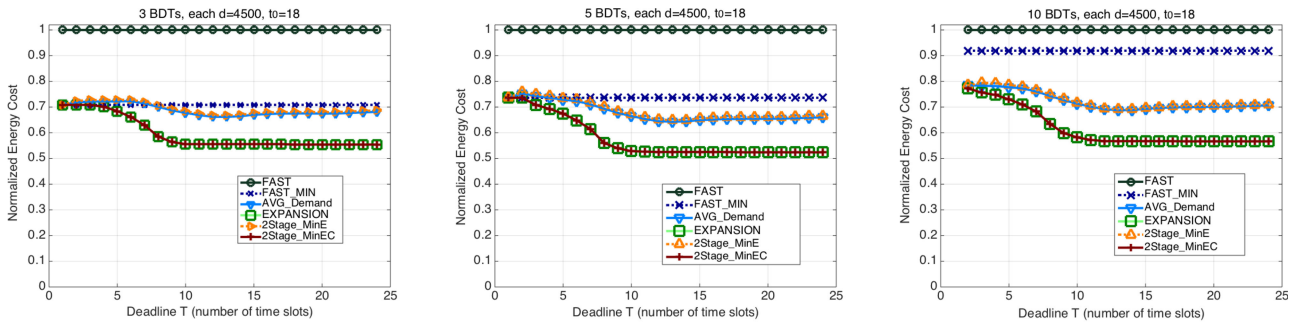


Fig. 8. Normalized energy-cost for all the comparison methods when $d = 4500, t_0 = 18$, and the number of BDTs is 3, 5, and 10 separately.

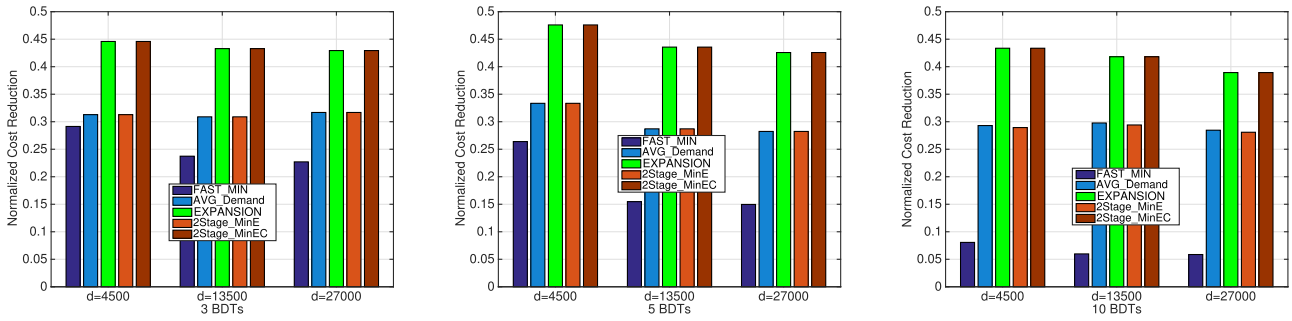


Fig. 9. Energy-cost reduction ratios for all the comparison methods with different number of BDTs when $T = 24$ and $t_0 = 18$.

Comparing Figs. 4 and 8, we can see the energy-cost of our method is further reduced (the ratio reaches 48 percent) when $t_0 = 18$. This means our method gains more benefits on energy-cost savings when the scheduling start time t_0 is in the busy hour. Such further reduction is also confirmed by comparing Figs. 6 and 9. Both of them show the energy-cost reduction ratios when $T = 24$ for different combinations of demand d and the number of inter-DC BDTs. Fig. 9 also verifies the aforementioned trend of our method on the data transfer volume. Note for that when $t_0 = 18$, the energy-cost of the AVG_Demand and 2Stage_MinE methods show a different variation trend compared with the case of $t_0 = 1$. The energy-cost of them increases first (just a few time slots) and then decreases when T varies from 1 to 24. This is also due to the real-life electricity price variations, as described before (after the busy hour, the electricity price generally shows a decreasing trend).

We also test all the methods when t_0 is in the other time slots. The results show that the performance of our method on energy-cost savings is between the free ($t_0 = 1$) and the busy hour ($t_0 = 18$), but it always achieves the minimum energy-cost in all the methods. We skip the details due to the limit of space.

5.3.5 Large-Scale Simulation

Since the scale of real-life inter-DC network is limited to the actual reality.[2] Next, we will evaluate the performance of our method with large-scale simulations. In this simulation, we use the random graph based algorithm to generate inter-DC network topologies with different number of nodes. Each node (data center) resides in a randomly selected regional electricity market. If more than one data center locates in the same electricity market, we use different hubs to differentiate them. The value of other parameters is set the same as before.

First, we let the number of data centers varies from 10 to 100. For each inter-DC network, we fix the number of inter-DC BDTs to half of the number of the nodes, $d = 4500$, $t_0 = 18$ (busy hour), and $T = 24$. Note for that the EXPANSION method is not included in this simulation, since it is too slow to solve the optimization solution in the large-scale environments. Fig. 10 illustrates the energy-cost reduction ratios compared with the FAST method. As it shows, our method (2Stage_MinEC) achieves the largest energy-cost reduction ratios in all the cases. The energy-cost reduction

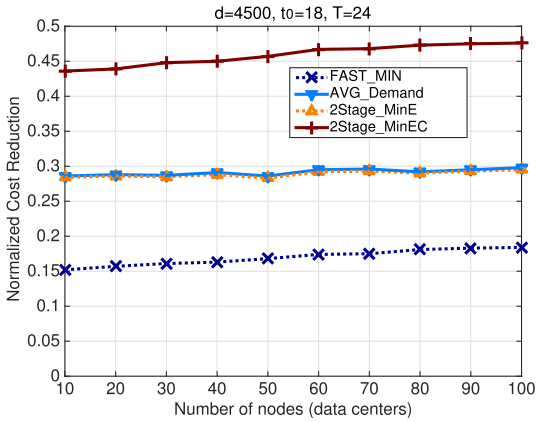


Fig. 10. Normalized cost reduction ratios with different number of nodes.

2. Due to commercial concerns, we failed to find out a larger real-life inter-DC network.

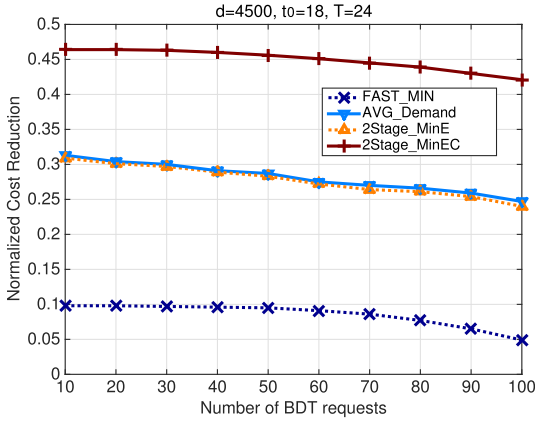


Fig. 11. Normalized cost reduction ratios with different number of BDT requests.

ratios of our method are much higher than that of any other methods. It demonstrates that significant energy-cost savings of our method can be achieved in the large-scale environment too. Furthermore, the energy-cost savings of our method do not deteriorate as the number of nodes increases. It is a desirable property, especially when the scale of inter-DC networks may increase significantly in the future. The main reason is that for larger scale of inter-DC networks, more geo-distributed data centers (may with lower electricity prices) can be used to perform the inter-DC BDTs, which brings more temporal and spatial flexibility to minimize the energy-cost.

Then, we fix the number of nodes to 50, and vary the number of inter-DC BDTs to evaluate these methods on different scales of data transfer volumes. As Fig. 11 shows, the energy-cost reduction ratio of our method is always larger than that of any other methods when the number of inter-DC BDTs varies from 10 to 100. It shows that our method can bring significant energy-cost savings over existing BDT methods, even when the scale of inter-DC BDTs is large. As the number of inter-DC BDTs increases, the energy-cost reduction ratios of FAST_MIN method decreases rapidly than other three methods. When the number of inter-DC BDTs is 100, only achieves 5 percent cost reduction. Though our method shows a slight decreasing trend as the number of inter-DC BDTs increases, it still outperforms other methods on energy-cost savings. This demonstrates that significant energy-cost savings of our method can be achieved, even when the scale of inter-DC BDTs is constantly expanding.

To conclude, above evaluations with real-life inter-DC network and real-life electricity prices show that our method brings significant energy-cost savings over existing methods for inter-DC BDTs. The large-scale simulation demonstrates that significant energy-cost savings of our method can be achieved not only in the real limited case study, but also in the large-scale simulation environment.

6 Conclusion

The fast proliferation of cloud computing promotes the rapid growth of large-scale commercial data centers. Geo-distributed data centers are often used by major cloud service providers to provide customers with better reliability and quality of service. In such large-scale geo-distributed data center networks, Inter-DC bulk data transfer becomes an important and increasing requirement, due to huge amounts of data (periodic data backup, software distribution, virtual machines cloning, distributed databases, etc.) need to be transferred between these data centers. Although geographic distribution and large delay-tolerance of inter-DC BDTs have been used by many works to reduce the operating cost of geo-distributed data centers, there are still a lot of problems remain unexplored. Motivated by that existing works for inter-DC bulk data transfers mainly focus on optimizing the bandwidth cost, we develop an efficient two-stage optimization method to solve the novel problem of minimizing the energy-cost for geo-distributed data centers in the multi-electricity-market environment. For each BDT, the proposed method first searches for the optimal demand division along available time slots, and then computes the optimal route and schedule for each time slot respectively. Extensive evaluations with real-life inter-DC network and real-life electricity prices show that the proposed two-stage optimization method brings significant energy-cost savings over existing inter-DC bulk data transfer scheduling methods.

In the future, we plan to integrate the store-and-forward capability of the intermediate data centers to further minimize the energy-cost for inter-DC BDTs, i.e., the intermediate data centers are able to temporarily store the bulk data to be relayed, and forward them at a later time to its downstream relay nodes or to the destination. By appropriately determine when and how much should the bulk data be stored at the intermediate data centers, more energy-cost savings could be achieved for inter-DC bulk data transfers.

Acknowledgments

This work was partially supported by the NSF of China under grant No. 61602175, the National Key Research and Development Plan under grant No. 2016YFA0502300, the National Science and Technology Supporting Program of China under grant No. 2015BAH18F02, the Model Information Service Industry Program of Guangdong Province under grunt No. GDEID2010IS049, the Fundamental Research Funds for the Central Universities under grant Nos. 222201514331 and ZH1726108, and the Special Funds for Informatization Development in Shanghai under grant No. 201602008.

References

[1] D. Xu and X. Liu, "Geographic trough filling for internet datacenters," in *Proc. IEEE INFOCOM*, 2012, pp. 2881–2885.

[2] Y. Wu, Z. Zhang, C. Wu, C. Guo, Z. Li, and F. C. M. Lau, "Orchestrating bulk data transfers across geo-distributed datacenters," *IEEE Trans. Cloud Comput.*, vol. 5, no. 1, pp. 112–125, Jan.–Mar. 2017.

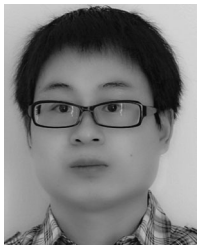
[3] N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, "Inter-Datacenter Bulk Transfers with NetStitcher," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 74–85, 2011.

[4] Y. Feng, B. Li, and B. Li, "Postcard: Minimizing costs on inter-datacenter traffic with store-and-forward," in *Proc. 32nd Int. Conf. Distrib. Comput. Syst. Workshops*, 2012, pp. 43–50.

[5] F. Research, "The future of data center wide-area networking," 2010. [Online]. Available: info.infineta.com/l/5622/2011-01-27/Y26

[6] F. Kong, X. Lu, M. Xia, X. Liu, and H. Guan, "Distributed optimal datacenter bandwidth allocation for dynamic adaptive video streaming," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 531–540.

[7] Y. Chen, S. Jain, V. K. Adhikari, Z.-L. Zhang, and K. Xu, "A first look at inter-data center traffic characteristics via Yahoo! datasets," in *Proc. IEEE INFOCOM*, 2011, pp. 1620–1628.
[8] T. Nandagopal and K. P. N. Puttaswamy, "Lowering inter-datacenter bandwidth costs via bulk data scheduling," in *Proc. 12th IEEE/ACM Int. Symp. Cluster Cloud Grid Comput.*, 2012, pp. 244–251.
[9] Y. Feng, B. Li, and B. Li, "Jetway: Minimizing costs on inter-datacenter video traffic," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 259–268.
[10] N. Laoutaris, G. Smaragdakis, R. Stanojevic, P. Rodriguez, and R. Sundaram, "Delay-tolerant bulk data transfers on the internet," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1852–1865, Dec. 2013.
[11] Y. Wang, S. Su, S. Jiang, Z. Zhang, and K. Shuang, "Optimal routing and bandwidth allocation for multiple inter-datacenter bulk data transfers," in *Proc. IEEE Int. Conf. Commun.*, 2012, pp. 5538–5542.
[12] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: Research problems in data center networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 68–73, 2008.
[13] I. Alan and T. Kosar, "Energy-aware HTTP data transfers," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. Workshops*, 2016, pp. 37–42.
[14] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, "Energy proportional datacenter networks," *ACM SIGARCH Comput. Archit. News*, vol. 38, no. 3, pp. 338–347, 2010.
[15] Y. Shang, D. Li, and M. Xu, "Energy-aware routing in data center network," in *Proc. 1st ACM SIGCOMM Workshop Green Netw.*, 2010, pp. 1–8.
[16] J. Kwak, O. Choi, S. Chong, and P. Mohapatra, "Dynamic speed scaling for energy minimization in delay-tolerant smartphone applications," in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 2292–2300.
[17] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright, "Power awareness in network design and routing," in *Proc. IEEE 27th Conf. Comput. Commun.*, 2008.
[18] I. Alan, E. Arslan, and T. Kosar, "Energy-aware data transfer algorithms," in *Proc. Int. Conf. High Performance Comput. Netw. Storage Anal.*, 2015, Art. no. 44.
[19] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 123–134, 2009.
[20] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment," in *Proc. IEEE INFOCOM*, 2010, pp. 1–9.
[21] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely, "Data centers power reduction: A two time scale approach for delay tolerant workloads," in *Proc. IEEE INFOCOM*, 2012, pp. 1431–1439.
[22] X. Lu, F. Kong, J. Yin, X. Liu, H. Yu, and G. Fan, "Geographical job scheduling in data centers with heterogeneous demands and servers," in *Proc. IEEE Int. Conf. Cloud Comput.*, 2015, pp. 413–420.
[23] N. Laoutaris, G. Smaragdakis, P. Rodriguez, and R. Sundaram, "Delay tolerant bulk data transfers on the internet," *SIGMETRICS Performance Eval. Rev.*, vol. 37, no. 1, pp. 229–238, Jun. 2009.
[24] J. Shuja, et al., "Survey of techniques and architectures for designing energy-efficient data centers," *IEEE Syst. J.*, vol. 10, no. 2, pp. 507–519, Jun. 2016.
[25] X. Liu and F. Kong, "Datacenter power management in smart grids," *Found. Trends® Electron. Des. Autom.*, vol. 9, no. 1, pp. 1–98, 2015.
[26] A. Wierman, L. L. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems," in *Proc. IEEE INFOCOM*, 2009, pp. 2007–2015.
[27] J. Shuja, K. Bilal, S. A. Madani, and S. U. Khan, "Data center energy efficient resource scheduling," *Cluster Comput.*, vol. 17, no. 4, pp. 1265–1277, 2014.
[28] M. Lin, A. Wierman, L. L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1378–1391, Oct. 2013.
[29] J. Shuja, A. Gani, S. Shamshirband, R. W. Ahmad, K. Bilal, and L. Kazmerski, "Sustainable cloud data centers: A survey of enabling techniques and technologies," *Renewable Sustainable Energy Rev.*, vol. 62, pp. 195–214, 2016.
[30] F. Kong and X. Liu, "GreenPlanning: Optimal energy source selection and capacity planning for green datacenters," in *Proc. ACM/IEEE 7th Int. Conf. Cyber-Phys. Syst.*, 2016, pp. 1–10.
[31] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "GUPT: Privacy preserving data analysis made easy," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 349–360.
[32] A. Vulimiri, C. Curino, E. Godfrey, and P. Brighten, "Global analytics in the face of bandwidth and regulatory constraints," in *Proc. 12th USENIX Conf. Netw. Syst. Des. Implementation*, 2015, pp. 323–336.
[33] IshfaqAhmad and SanjayRanka, *Handbook of Energy-Aware and Green Computing*. Boca Raton, FL, USA: CRC Press, 2012.
[34] G. Prekas, M. Primorac, A. Belay, C. Kozyrakis, and E. Bugnion, "Energy proportionality and workload consolidation for latency-critical applications," in *Proc. 6th ACM Symp. Cloud Comput.*, 2015, pp. 342–355.
[35] H. Xu and B. Li, "Joint request mapping and response routing for geo-distributed cloud services," in *Proc. IEEE INFOCOM*, 2013, pp. 854–862.
[36] S. Narayana, J. W. Jiang, J. Rexford, and M. Chiang, "To coordinate or not to coordinate? wide-area traffic management for data centers," Princeton Univ., Princeton, NJ, USA, Tech. Rep. TR-998-15, 2012.
[37] D. Niu, H. Xu, B. Li, and S. Zhao, "Quality-assured cloud bandwidth auto-scaling for video-on-demand applications," in *Proc. IEEE INFOCOM*, 2012, pp. 460–468.
[38] K. Papagiannaki, N. Taft, Z.-L. Zhang, and C. Diot, "Long-term forecasting of internet backbone traffic: Observations and initial models," in *Proc. IEEE INFOCOM*, 2003, pp. 1178–1188.
[39] S. U. Khan and A. Y. Zomaya, *Handbook on Data Centers*. New York, NY, USA: Springer, 2015.
[40] D. Lozovanu and D. Stratila, "Optimal flow in dynamic networks with nonlinear cost functions on edges," in *Analysis and Optimization of Differential Systems*. Boston, MA, USA: Springer pp. 247–258, 2003.
[41] T. Hasuike, H. Katagiri, H. Tsubaki, and H. Tsuda, "Route planning problem with groups of sightseeing sites classified by tourist's sensitivity under time-expanded network," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, Oct. 2014, pp. 188–193.
[42] Softlayer datacenters map. [Online]. Available: http://www.softlayer.com/advantages/network- overview/
[43] Federal energy regulatory commission. [Online]. Available: http://www.ferc.gov/market-oversight/mkt-electric/overview.asp



Xingjian Lu received the PhD degree from Zhejiang University, China, in 2014. He is an assistant professor in the School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China. His current research interests include cloud computing, data center, energy management, performance evaluation, and optimal scheduling for different kinds of workloads. He has published more than 20 research papers in major peer-reviewed international journals and conference proceedings in these areas. E-mail: lulxj@ecust.edu.cn.

Fanxin Kong received the PhD degree in computer science from McGill University. He is a postdoctoral researcher in the Department of Computer & Information Science, University of Pennsylvania. His research interests include security, sustainability and timing aspects for cyber-physical systems, and Internet of Things. His research spans application areas of automobiles, transportation systems, energy systems, and data centers. He develops and applies techniques from game theory, distributed computing, machine learning, data analysis, deterministic and stochastic optimization, and approximate and online algorithm design for these areas. He has published more than 30 research papers in major peer-reviewed international journals and conference proceedings in these areas.

Xue Liu received the BSc and master's degrees from Tsinghua University, and the PhD (with multiple Hons.) degree in computer science from the University of Illinois at Urbana-Champaign. He is a William Dawson scholar and full professor in the School of Computer Science, McGill University. He has also worked as the Samuel R. Thompson Chaired associate professor with the University of Nebraska-Lincoln and as a visiting scientist at HP Labs, Palo Alto, California. His research interests are in computer and communication networks, real-time and embedded systems, cyber-physical systems and IOT, green computing, and smart energy technologies. He has published more than 200 research papers in major peer-reviewed international journals and conference proceedings in these areas and received several best paper awards. His research has been covered by various news media reports. He is a member of the IEEE.

Jianwei Yin received the PhD degree from Zhejiang University, China, in 2001. He is a full professor in the College of Computer Science and Technology, Zhejiang University, China. His current research interests include cloud computing, performance evaluation, service computing, middleware etc. He has published more than 120 research papers in major peer-reviewed international journals and conference proceedings in these areas. Furthermore, he has received more than 30 patents in the past five years. E-mail: zjuyjw@zju.edu.cn.

Qiao Xiang received the Master and PhD Degrees in Computer Science from Wayne State University, United States, in 2012 and 2014, respectively. He received his Bachelor Degrees in Engineering and in Economics from Nankai University, China, in 2007. He is a Postdoctoral Fellow with the Department of Computer Science at Yale University, United States. Before that, he was a Postdoctoral Fellow with the School of Computer Science at McGill University, Canada. His research interests include software defined networking, data center networks, cyber physical systems, vehicular networks, smart grid and network economics. E-mail: qiao.xiang@cs.yale.edu.

Huiqun Yu received the BS degree from Nanjing University, in 1989, the MS degree from East China University of Science and Technology (ECUST), in 1992, and the PhD degree from Shanghai Jiaotong University, in 1995, all in computer science. He is currently a professor of computer science in the Department of Computer Science and Engineering, ECUST. From 2001 to 2004, he was a visiting researcher in the School of Computer Science, Florida International University. His research interests include software engineering, high confidence computing systems, cloud computing, and formal methods. He is a senior member of the IEEE, a member of the ACM, and a senior member of the China Computer Federation. E-mail: yhq@ecust.edu.cn.