IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 14, 2017, accepted November 28, 2017, date of publication December 8, 2017,
date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2781309

EC^3: Cutting Cooling Energy Consumption Through Weather-Aware Geo-Scheduling Across Multiple Datacenters

JIAHONG WU, YUAN JIN, AND JIANGUO YAO, (Senior Member, IEEE)
School of Software, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Jianguo Yao (jianguo.yao@sjtu.edu.cn)

This work was supported in part by the Program for NSFC under Grant 61772339 and Grant 61572322, in part by the STCSM Project under Grant 16QA1402200, in part by the MOST under Grant 2013FY111900, and in part by Shanghai Jiao Tong University under Grant YG2015MS47.

ABSTRACT Most information technology (IT) equipment found in a data center is air-cooled as electrical component produces heat, which must be removed to prevent the temperature of the IT equipment from rising to an unacceptable level. The energy consumption for the data center cooling system is positively related to the air temperature outside the data center. The difference of data center internal temperature and the outside air temperature varies from each data center location. If we reschedule the workload of Internet cloud services to the least temperature difference, the cooling energy consumption will be the biggest savings. The cooling energy-consumption model and query characteristics of cloud services provide the methodology to formulate the energy consumption and workload rescheduling. However, the cloud service must meet the tail latency constraint after the rescheduling. We solve this problem by estimating the high-percentile tail latency and scheduling the cloud service to where can meet the tail latency constraint. At last, a proactive weather-aware geo-scheduling algorithm, called EC^3, is proposed to distribute end-users' loads among data centers so as to reduce the cooling energy consumption. The trace-driven experiments on real clouds and data center workload traces show the effectiveness of our design for reducing data center cooling consumption.

INDEX TERMS Cooling energy, weather-aware, data center, internet cloud service.

I. INTRODUCTION

The data centers for cloud computing has evolved significantly during the past decades by adopting more efficient technologies and practices in data center infrastructure management. Current study results show the U.S energy consumption of data centers has increased dramatically, accounting for about 1.8% of the U.S. electricity usage in 2014 and the data center electricity consumption increased by about 4% from 2010-2014. Energy use is expected to continue slightly rising in the near future, increasing 4% from 2014-2020, the same rate as the past five years. Based on current trend estimates, U.S. data centers are projected to consume approximately 73 billion kWh in 2020 [1].

The cooling energy consumption is an important issue for minimizing environmental impact, lowering costs of energy consumption and optimizing data center operation performance. A modern data center has a large room with many rows of racks filled with a huge number of servers and other information technology (IT) equipment used for processing, storing and transmitting digital information, and an amount of heat is generated by the IT equipment. To maintain the reliability of the IT equipment in the data center, it is of importance to maintain proper temperature. This work [2] presents the results of an investigation of 10 random data centers. It reveals the representative energy consumption distribution and the variation, showing a spread between 30% and 55% of the total energy consumed by cooling the data center. Cooling and ventilation systems consume on average about 50% of the total energy used. Therefore, how to reduce the cooling energy consumption of data center must be taken into account.

The cooling energy consumption of the data center is more complicated than the conventional IT equipment since the energy usage of a cooling system varies with the outside temperature. Essentially, the hotter the outside temperature, the more energy it takes to chill a data center. The outside temperature does not just vary from one location to another, it varies all the time, and any single day is usually at least a little bit warmer or colder than the day before it. If we use the cooling system to keep the data center at a roughly constant temperature, the amount of energy that the cooling system uses will vary from one day to the next. Cooling Degree Days (CDD), are a measure of how much (in degrees), and for how long (in days), outside air temperature was higher than a specific base temperature. They are used for calculations relating to the energy consumption required to cool buildings [3]. The number of degrees that the average air temperature is above 23^{o}C and the cooling system starts to cool the building. To calculate the CDD, take the average of day's hour high and low and subtract 23. The cost of weather derivatives trading is based on an index made up of CDD values. The settlement cost for a weather futures contract is calculated by summing the CDD values and multiplying by unit costs (such as 20).

Cloud Service Provider (CSP) has deployed multiple data centers worldwide [4]. To save cooling energy consumption, the workloads can change to running in the data center with cool weather. However, there are other factors that we must consider before the workload rescheduling, such as how many workloads should be rescheduled, and whether the Cloud Services meet the tail latency constraint after the rescheduling. Therefore, we are struggling to solve this complex problem in this paper. Our objective is to optimize the workload distribution such that the data center cooling energy consumption is reduced. The temperature difference in different data centers vary widely, and it is related to the local weather. So, the workload rescheduling balances end-users' load based on the data center outside air temperature to save cooling energy consumption. We present a comprehensive energy consumption model to formulate the workload rescheduling. Through the mathematical model and a novel data-driven latency estimation approach, we propose an algorithm which is weather aware for the dynamic workload rescheduling among multiple data centers.

To summarize, our contributions in this paper include:

- We develop a cooling energy consumption model. It considers the energy use to reduce the heat conducted with outside world and the heat generated by the IT equipment. To do this, we follow the CDD to calculate temperature difference and formulate the cooling energy consumption. In particular, we construct an optimization function by binding the high-percentile tail latency with SLA [5].
- We novelty propose EC^3 to make scheduling decisions for the Cloud Services. It dynamically schedules the workloads to data center with least internal-outside temperature difference while meeting the tail latency SLA requirements. To our best knowledge, our work is the first that takes a holistic approach by covering the cooling energy consumption of the data centers from CDD to adjust the workloads scheduling in time slot.
- Under a variety of settings, we conduct extensive simulation-based experiments using real clouds and data center workload traces. The latency estimation is based on the historical Cloud Service's request. The estimation of cooling energy consumption is achieved by the correlation of data center energy bill and CDD. The experimental result turns out that with workload rescheduling the cooling consumption is reduced more than 40% while Cloud Service can still meet the tail latency SLA constraint.

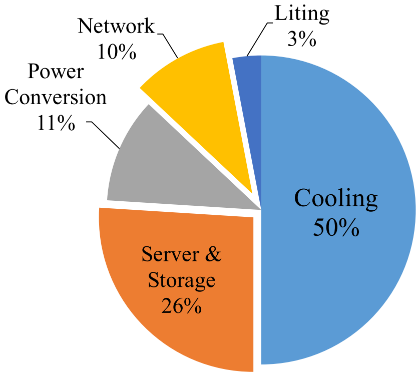


FIGURE 1. A breakdown of energy consumption by different components of a data center [9].

II. BACKGROUND AND MOTIVATION

A. BACKGROUND

Data center contains large numbers of compute nodes to support the increasing workloads and provide promising Quality of Service (QoS) for the Cloud Services. Besides, electronic components are continually becoming smaller and more powerful, data centers must deal with the heat generated by having a large number of high-power processors tightly packed into a small space. Increases in energy consumption usually come in double-helpings in cloud computing, since the amount of energy required to cool an object is theoretically increasing as it consumes more energy to support more workloads. The energy consumption by different components of a data center is shown in Fig. 1. The power of the ancillary facilities used to maintain the normal operation of the data center is rising. Up to 50% of total energy utilized for the data center is wasted for cooling the data center [6].

The problem of exponentially-growing cooling energy consumption becomes even more pronounced when considering global warming [7]. It has reached a point where the energy required to chill the data centers is a greater financial burden than the hardware and maintenance costs put together [8]. So it would seem reasonable that CSPs can dramatically cut overall energy consumption by creatively eliminating the data center cooling energy consumptions.

B. MOTIVATION

One solution to save cooling energy consumption would be to build new data centers in the area that has a cold climate. This solution is now the direction that the cloud computing industry is headed. Google recently built a major $230,000,000 data center in Finland, hoping to leverage the

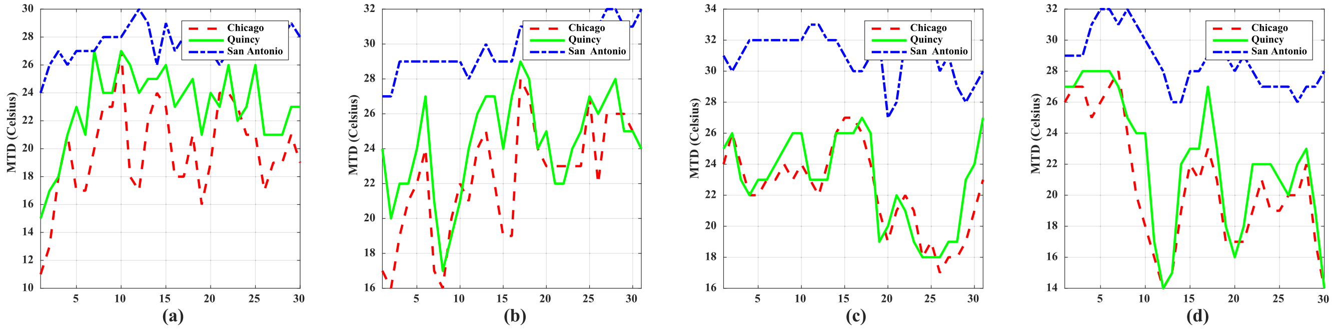


FIGURE 2. Mean temperature of days of three Microsoft data center locations in four months. (a) Apr; (b) May; (c) Jun; (d) Jul.

cold temperatures to lower the cooling energy consumption. Facebook also constructed a new data center near the Arctic Circle in Lulea, Sweden, because of the region's cold weather [10]. This move towards colder regions will help dramatically cut cooling energy consumption associated with heating.

The other solution is when the CSP has built data centers, and inside the data center are server utilization-based smart temperature monitoring [11] and sub-system cooling [12]. For example, an independent cooling system chills only one sub-cluster and each independent cooling system shutdowns automatically when the sub-cluster has maintained at target temperature or is unused. As the data center temperature needs to be maintained at around 23°C, the cooling costs are related to the data center load and the outside temperature [13]. The good news is the replication of data sets among data centers provides an opportunity to reschedule the Cloud Services. If the data sets are not replicated, the energy used for data migration may be more than the cooling energy savings.

In fact, Microsoft has built data centers in Chicago, Quincy and San Antonio. Fig. 2 shows the mean temperature day of four months for these cities [3]. The workloads can be rescheduled to less internal-outside temperature difference data center, such as from San Antonio to Chicago. The data center located in San Antonio shutdowns the cooling machine for unused clusters to save cooling energy consumption. There are some other constraints that we must consider. Some Cloud Service requires stringent tail latency bounds; different Cloud Services have varying levels of data consistency; some CSP prefers those location closed data center which can quickly provide services. All of these requirements result in a nonconvex optimization with no efficient solution. Hence, we propose an efficient and greedy heuristic that dynamically places workloads to "better" data centers. It greedily maximizes the expected reduction in cooling energy consumption normalized by moving the workloads to least internal-outside temperature difference data center and closing the unnecessary cooling system in the original data center.

III. RELATED WORK

It can be presumed that there is a lot of potential for energy conservation strategies when providing cooling to data centers, given the weather conditions in climate zones. However, the multi-potential of variable cooling solutions has not been studied in the previous literature.

Tang *et al.* [14] introduce three heuristic approaches to minimize the total energy of a data center by scheduling workloads to have a uniform outlet temperature profile, minimum server power dissipation, or a uniform workload distribution, respectively. This work [15] is to maximize reward rate, which mitigates the impact of co-location interference by maximizing a reward rate objective function that considers co-location interference. It maximizes the reward rate earned by the system while obeying red-line temperature thresholds and a power constraint on the whole facility (both compute and cooling power). However, these works do not consider the tail latency constraint of Cloud Services.

This work [16] focuses on energy minimization in a data center accounting for both the IT equipment and the cooling power usage. In particular, they address the server consolidation concurrently with the task assignment. However, authors do not consider the temperature difference of the data center locations. This work [17] propose a unified management approach with leveraging stochastic optimization tools which allow the data centers to adaptively respond to a variability of cooling efficiency under long-term QoS requirements. However, the proposed algorithm yields a strategy without knowing the distributions of the independently and identically distributed.

There are other techniques currently employed to reduce the energy cost and power density in data centers. For example, load balancing [18], [19] can be used to distribute the workloads of the data center among different servers evenly to balance the per server workload (and hence achieve uniform power density). Server consolidation [20], [21], which refers to assigning incoming tasks to the minimum number of active VMs in the data center and shutting down unused VMs, is another approach for power reduction of data centers. Also, there have been many recent works that consider thermal-aware resource management (e.g., [22]–[26]). Some do not consider heterogeneity in their work [22], [23], [26]. Others do not consider Dynamic Voltage and Frequency Scaling (DVFS) control [24], [25].

IV. SYSTEM MODEL

A. ENERGY CONSUMPTION MODEL

1) HEAT CONDUCTION

We know that a closed space will not be heat conduction with the outside world, and the air temperature of this closed space thus unchanged. We at first do not consider the heat generated by the data center IT equipment (e.g., servers, switch), and the temperature rise because of the heat conduction with outside world. CDD provides a productive way to quantify the cooling energy consumption for reducing the heat conducted with the outside world. The idea is the amount of energy needed in any day is directly proportional to the number of CDD in that day [27]. With this method, we just need to do the linear regression analysis in each data center and correlate the energy consumption data with CDD. Besides, this tool [3] enables us to access CDD data on a variety of timescales. We can get energy bills from a utility or energy supplier. Although the occupancy of the data center and the cooling patterns might vary throughout in every time slot, the patterns are usually fairly consistent from one time slot to the next.

With the two data sets (energy consumption record and CDD), we can use common tools, such as Matlab, to do the linear regression analysis. Let $f(c)$ denote the estimated linear regression function and parameter c is the CDD for the specific data center and specific date. The result of $f(c)$ is the amount of estimated energy needed to cool a data center without running IT equipment.

$$f(c) = a * c + b. \tag{1}$$

2) HEAT PRODUCTION

To calculate the direct costs associated with running IT equipment, we need to know the direct energy consumption of each type of IT equipment and the costs associated with cooling the environment where the IT equipment is situated. Let W and B denote the watt-hour and the British Thermal Unit (BTU) of one specific configuration IT equipment as this equipment in full utilization. In particular, BTU is used as a unit for the power of an air conditioning system and refers to the amount of thermal energy removed from an area. A BTU is approximately a third of one watt-hour [28].

Assume we calculate the total energy consumption by running one IT equipment for h hours (called the operating hour). The total energy consumption can be split into two parts, one for running this equipment and second the energy for reducing the heat produced by this equipment. Let C_1 and C_2 denote the energy consumptions for in-service use and cooling, respectively.

Energy Consumption for In-Service Use: For calculating the energy consumption and associated costs, we use the following:

$$C_1 = h * W/1000. \tag{2}$$

Note that the division operation on 1000 is for the conversion of Wh to kWh.

Energy Consumption for Cooling: For calculating the cooling consumption to keep the IT equipment in normal operating conditions, we use Eq. (3) where we translate the BTU to watt-hour, so we get the energy consumption associated with reducing the heat the IT equipment emits for h hours [28].

$$C_2 = \frac{h * B * 0.293}{1000 * \mathrm{COP}}. \tag{3}$$

Note that 1000 BTU is approximately 293Wh. COP is the ratio of useful output to the amount of energy input, used generally as a measure of the energy-efficiency of cooling or heating devices. COP equals heat delivered in BTU per hour divided by the heat equivalent of the energy input. Higher the COP, higher the efficiency of the devices [29].

3) TOTAL ENERGY CONSUMPTION

We can get the energy consumption associated with specific CDD by summing the cooling for heat conduction, the in-service use IT equipment and the corresponding cooling for produced heat. The linear function shown below defines the calculation.

$$e_j\left(a_j\right) = f(c) + \sum\nolimits_{k=1}^{K}\left(C_1^k + C_2^k\right). \tag{4}$$

Note that $e_j\left(a_j\right)$ is the energy consumption of data center j when total a_j workloads placed in data center $j \cdot f(c)$ is the correlation function of the cooling energy consumption and CDD, shown as Eq. (1). K denotes the total IT equipment of data center j.

B. WORKLOAD DISTRIBUTION MODEL

1) CLOUD SERVICE

Geo-distributed Cloud Services need data centers from different regions to accomplish data analysis workloads with lower tail latency [30]–[33]. A key novelty of our study is that each Cloud Service request is simultaneously sent to one data center. That is, each Cloud Service needs a group of data centers (called data center group) to complete the Cloud Service's workloads. For distributed services, non-distributed services are a special case which is hosted in a single data center, so our study can meet the query characteristics of these Cloud Services.

We assume that one CSP has a set N of data centers. Thus, there are maximum $|N|$ possible data centers for workloads of the Cloud Services. Note that this assumes data replication across the data centers. It is subject to data residency requirement. If data is not replicated, we consider each un-replicated data center as one data center group. The running workload depends on the data set [30], and the data replication provides the possibility of the workload migration.

Assume there are S different traffic sources. g is the set of data centers that can accept requests from the traffic source i. Note that not all the N data centers can accept request from traffic source i because of the data set replication constraint. Thus, we have a workload distribution decision vector $\vec{\lambda}_i = \{\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ig}\}^T$, where $\lambda_{ij} \geq 0$ denotes the amount of requests sent to data center j from traffic source i, and $\Lambda_i = \sum_{j=1}^{g} \lambda_{ij}$ is the total requests from traffic source i. Hence, the total workload sent to data center j can be expressed as:

$$a_j = \sum_{i=1}^{S} \lambda_{ij}. \tag{5}$$

The workload distribution decision now is to determine the load distribution $\vec{\lambda}_i$, for all traffic sources. Considering all the traffic sources, we define the load distribution matrix $\vec{\lambda} = \{\vec{\lambda}_1, \vec{\lambda}_2, \ldots, \vec{\lambda}_S\}$, which is the main decision variable in the optimization problem. Thus, the problem of deciding the workload distribution to each data center generalizes the existing global load balance literature [18], [34] that only decides workload distribution to each single data center.

2) TAIL LATENCY

We now consider the latency performance constraint between the traffic source and the data center. In this paper, the front-end gateway denotes the concentrated traffic source of Cloud Service requests. Specifically, the high-percentile latency of requests originating from each traffic source must be no greater than the corresponding threshold. For example, if x% is the percentile latency requirement, then at least x% of the requests must have latency not exceeding the threshold.

A key challenge is how to determine the tail latency for each Cloud Service. We need to examine each route/path between a traffic source and a data center. Since we have S traffic sources and N data centers, there is a maximum of $R = S \times N$ routes, each representing a network path from a traffic source location to a data center location. We account for the route from traffic source i to data center j by r_{ij}. In this paper, we primarily focus on data center-level workload scheduling decisions, while treating the scheduling decisions within each data center as irrelevant decisions. As such, the decision under consideration that affects a data center latency is equivalently the total amount of workloads sent to this data center. Hence, let $p_{ij}^{route}(a_j, r_{ij})$ denote the probability that the latency is less than D_i for route r_{ij}, given workload a_j at data center j.

Each Cloud Service needs to be processed in a group of data centers. We observe that the latency of Internet requests sent along one route is independent of that along with another route. The reason is that each Cloud Service request is small which taking no more than a few seconds to complete. These facts, combined with performance interference from other workloads, lead to the consequence that the tail latencies incurred in different data center group can be viewed as uncorrelated and independent. So, we combine the response time probabilities along different routes to express $\lambda_{ij} * p_{ij}^{route}(a_j, r_{ij})$ for requests from traffic source i to data center j. Further, since requests from traffic source i are distributed among the data center groups, the latency probability for requests from traffic source i should be averaged across the data center groups.

$$p_i(\vec{\lambda}) = \frac{1}{\Lambda_i} * \sum_{j=1}^{g} \lambda_{ij} * p_{ij}^{route}(a_j, r_{ij}). \tag{6}$$

We use $p_i(\vec{\lambda})$ to emphasize that the latency threshold satisfaction probability is the critical function of the workload distribution decision.

V. THE DESIGN OF EC^3

We now present the design of the dynamic weather-aware scheduling architecture, which reduces cooling consumption for data centers with a tail latency SLA constraint. First, we will define the formulation to account for workload distribution in Section V-A. Then, we discuss the latency profiling technique in Section V-B. At last, we outline the overview of EC^3 in Section V-C.

A. PROBLEM FORMULATION

Mathematically, the operator has the following workload rescheduling optimization function to save cooling energy consumption:

$$\text{Minimize } \sum_{j=1}^{N} e_j\left(a_j\right) \tag{7}$$

$$\text{subject to } p_i\left(\vec{\lambda}\right) \geq P_i^{SLA}, \quad \forall i \in 1, 2, \ldots, S \tag{7a}$$

$$\sum_{j=1}^{g} \lambda_{ij} = \Lambda_i, \quad \forall i \in 1, 2, \ldots, S \tag{7b}$$

$$a_j \leq U_j, \quad \forall j \in N. \tag{7c}$$

The optimization function (7) is the total energy consumption across all the N data centers. The constraint (7a) expresses the tail latency performance constraint set by the SLA, and P_i^{SLA} is the tail latency SLA requirement for traffic source i. The constraint (7b) ensures that all the Internet requests from each traffic source can be processed. The constraint (7c) ensures that the total workloads sent to a data center must not exceed the data center capacity.

B. LATENCY PROFILING

Up to this point, we have decomposed $p_i(\vec{\lambda})$, the probability that the tail latency is less than the threshold for all Internet requests from traffic source i, into Eq. (6). In order to solve this optimization function (7), we still need to determine $p_{ij}^{route}(a_j, r_{ij})$, the probability that the latency for requests along the route from traffic source i to data center j is less than the corresponding latency threshold D_i.

There is a lot of research results in the context of queuing performance in high load data centers (e.g., [18], [35] and the references therein). In particular, a central limit theorem for heavy traffic queuing systems states that for a G/G/m queue under heavy traffic load, the waiting time distribution could be approximated by an exponential distribution.

This theorem applies to the tail latency distribution as well, since the tail latency distribution converges to the waiting time distribution as the requests send to data center increases. The intuition behind this approximation is that in the high load data center, the long queuing effect helps effectively smooth out processing time fluctuations (i.e., the law of large numbers), which causes the waiting time or latency to converge to a distribution closely surrounding its mean value, i.e., the short-tailed exponential distribution, regardless of the actual arrival process and service time distribution. Inspired by this result, we further postulate that for one Cloud Service mapped to one data center group, the response time distribution $F_{GE}(x)$ for any arrival process can be approximated as a Generalized Exponential distribution function [36], [37], as follows,

$$F_{GE}(x) = \begin{cases} (1-e^{-\mu x})^{\alpha} & x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Note that μ and α are the scale and shape parameter, respectively. The mean $E(X)$ and variance $V(X)$ of the Cloud Service response time are given by [36], [37]

$$E(X) = \frac{1}{\mu}\left[\psi(\alpha+1)-\psi(1)\right], \tag{9}$$

$$V(X) = \frac{1}{\mu^2}\left[\psi'(1)-\psi'(\alpha-1)\right]. \tag{10}$$

Note that $\psi(.)$ and its derivatives are the digamma and poly-gamma functions.

From Eq. (9) and Eq. (10), we can know that the distribution in Eq. (8) is completely determined by the mean and variance of the historical requests' latency. The reason behind the use of this distribution, instead of the exponential distribution, is that it can capture both heavy-tailed and short-tailed task behaviors depending on the parameter settings and meanwhile, it degenerates to the exponential distribution at $\alpha = 1$ and $E(X) = 1/\mu$. This distribution significantly outperforms the exponential distribution in terms of tail latency predictive power for all the cases studied.

The implication of the latency estimation is significant. It allows not only the tail latency performance of a Cloud Service mapped to a diverse range of the data centers to be captured by a unified distribution function, but also the latency distribution of request and hence the tail latency SLA for the entire task-partitioning-merging system to be derived. We know that with all the requests of one Cloud Service send to data center group g being viewed as black boxes [37], one effectively transforms the task-partitioning-merging problem into a split-and-merge model whose distribution function can be expressed as $F_g(x)$.

$$F_g(x) = \begin{cases} \prod_{k=1}^{g}(1-e^{-\mu_k x})^{\alpha_k} & x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

It assumes the Cloud Service response times for requests mapped to different data centers are independent random variables. Now assume the parallel data centers are homogeneous [37], the distribution function can be further simplified as:

$$F_g(x) = \begin{cases} (1-e^{-\mu x})^{g\alpha} & x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

With Eq. (12), it can be easily realized that the p-th percentile request response time D_i for traffic source i can be written as [37]:

$$D_i = -\frac{1}{\mu}\log\left(1-\left(\frac{p}{100}\right)^{\frac{1}{g\alpha}}\right). \tag{13}$$

We now consider traffic source i sends W requests to data center j during each time slot. We define the method of estimating the high-percentile SLA tail latency as the following definition: The probability of request does not exceed the latency threshold can be calculated by the function shown below.

$$p_{ij}^{route} = \frac{\sum_{w}^{W}[d_{ij} \leq D_{ij}]_w}{|W|}. \tag{14}$$

Note that d_i is the expected/measured latency of each request. Note that operation $[d_i \leq D_i]$ represents a statistical method, and when $d_i \leq D_i$ the result is 1, otherwise 0. According to the law of large numbers, the average of the latency obtained from a large number of latency traces should be close to the expected value. The latency estimation will tend to become more accurate as more traces are considered in the estimation.

In Eq. (13), since D_i is a function of μ and α, which in turn are functions of mean and variance of the Cloud Service tail latency (according to Eq. (9) and Eq. (10)), a link between any given tail latency SLA, and $E(X)$ and $V(X)$ is established. The implication of this result is significant. On the one hand, with any given tail latency SLA, the resulting mean and variance can serve as the Cloud Service response time budgets for optimized workload distribution. On the other hand, with given measured Cloud Service tail latency statistics regarding mean and variance, we can predict whether the Cloud Service meets the target tail latency SLA with Eq. (6) and Eq. (14).

C. SYSTEM OVERVIEW

We show the overview of the weather-aware geo-scheduling architecture in Fig. 3. These are three main components in the optimizer: The cooling energy correlation component, the SLA tail latency estimation component and the GLB component. The inputs of the correlation component include the data center energy bill and CDD in each data center location. It outputs the cooling energy consumption for reducing the heat conducted with the outside world with specific CDD. The inputs of the SLA tail latency estimation component include the profiled tail latency of the Cloud Services and the estimated workload arrival at each traffic source during the current time slot. It outputs the percentile tail latency of each Cloud Service. Then the GLB component solves the optimization function (7) and outputs the optimized workload distribution decisions that split the incoming workloads at each traffic source to different data center groups to reduce the cooling energy consumption.

We present the general procedure for reducing cooling energy consumption as Algorithm 1. A key component of the proposed architecture is the latency profiler that determines the tail latency performance. It solves the problem using a

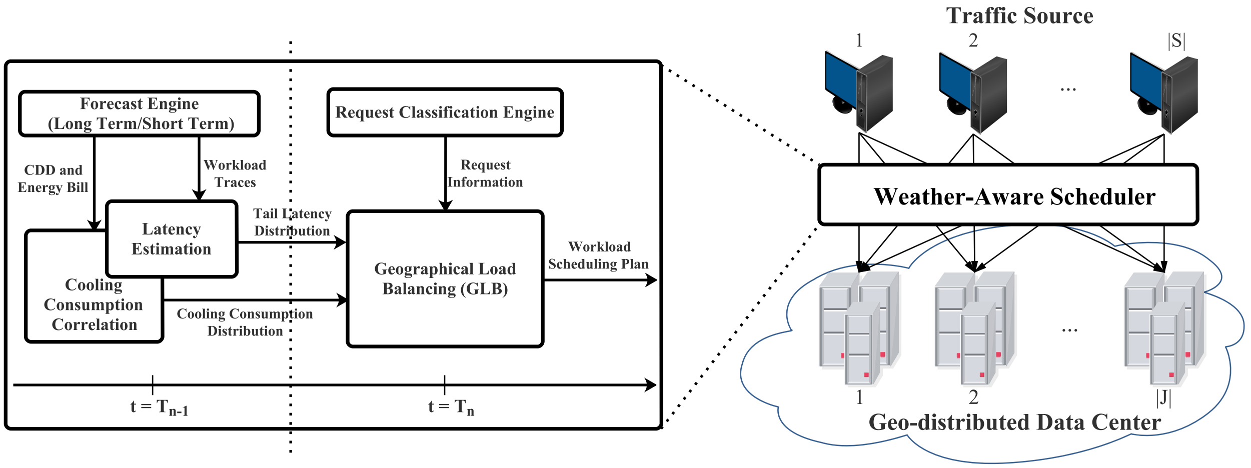

Forecast Engine (Long Term/Short Term)
CDD and Energy Bill
Workload Traces
Latency Estimation
Cooling Consumption Correlation
Tail Latency Distribution
Cooling Consumption Distribution
Request Classification Engine
Request Information
Geographical Load Balancing (GLB)
Workload Scheduling Plan
t = T$_{n-1}$
t = T$_n$
Traffic Source
1
2
...
|S|
Weather-Aware Scheduler
1
2
...
|J|
Geo-distributed Data Center

FIGURE 3. System overview of EC^3: The weather-aware geo-scheduling architecture for multiple data centers.

Algorithm 1 Procedure for Workload Rescheduling Algorithm EC^3

```
Step 1: For each data center location, initialize $f(c)$ with correlation of energy bill and CDD.
Step 2: Calculate $E(x)$ and $V(x)$ for each request in all possible data center running a scheduling policy, at desired request rate $\lambda_{ij}$;
Step 3: For each data center group, get $\alpha$ and $\mu$ with Eq. (8) by plugging in the measured $E(X)$ and $V(X)$ into Eq. (9) and Eq. (10), respectively;
Step 4: Estimate the SLA latency threshold $D_i$ based on Eq. (13).
Step 5: Choose the data center group recursively.
 Recursion($S$):
  $e$ = INF
  for $g$ in $G$ then:
    for $i$ in $S$ then:
      for $j$ in $g$ then:
        if $p_{ij}^{route} \le p_i^{SLA}$ and $a_j + i \le U_j$ then:
          $e_j(a_j) += ex$
        else
          $f = false$
          $break$
    if $f != false$ then:
      $e = \min(e,$ Recursion($next\ S) + e_j(a_j))$
  return $e$
 Note: $g$ denotes the candidate data center group. $ex$ denotes the added energy consumption with workload $i$ assigned to data center $j$.
```

numerical optimization method [38] and outputs the optimized workload distribution decisions.

VI. EVALUATION

In this section, we used real clouds and data center workload traces to evaluate the performance of EC^3. At first, we describe the experimental setup, in particular the real-world clouds, data center workload traces and CDD. Then, we explain how our latency estimation theorem works with real-world data center workload traces. At last, we evaluate the reduction of cooling energy consumption for multiple data centers.

A. SETTING

1) REAL-WORLD CLOUDS

We conducted some trace-driven experiments on data centers of Microsoft Azure [39]. These data centers are located in three places: Chicago, Quincy and San Antonio. The full power of the IT equipment, denoted as W, can be calculated from the product specifications. The BTU of each IT equipment can be calculated with conversion function $B = 0.293 * W$ [28].

2) DATA CENTER WORKLOADS AND LATENCY ESTIMATION

We consider the data center with workload arrival rates that can be predicted over a decision time slot [40]. Each Cloud Service includes several workloads which will be scheduled to different data centers. We take the data center workload traces as the basic trace-driven experimental data.

3) GETTING THE ENERGY CONSUMPTION DATA AND CDD

We get the records of energy consumption with energy bills from a utility or energy supplier. Most data centers follow a daily routine, which means that daily energy-consumption data is typically a good option for regression analysis. The occupancy of the data center and the cooling patterns are usually fairly consistent from one day to the next in the data center. Besides, the estimation of cooling energy consumption needs CDD of time slots. This tool [3] provides a way for us to get the CDD.

B. LINEAR REGRESSION ANALYSIS

We use the CDD and cooling energy consumption to plot an X-Y scatter chart of CDD against the consumption for the correlation. There are two notable extras that we can get from the

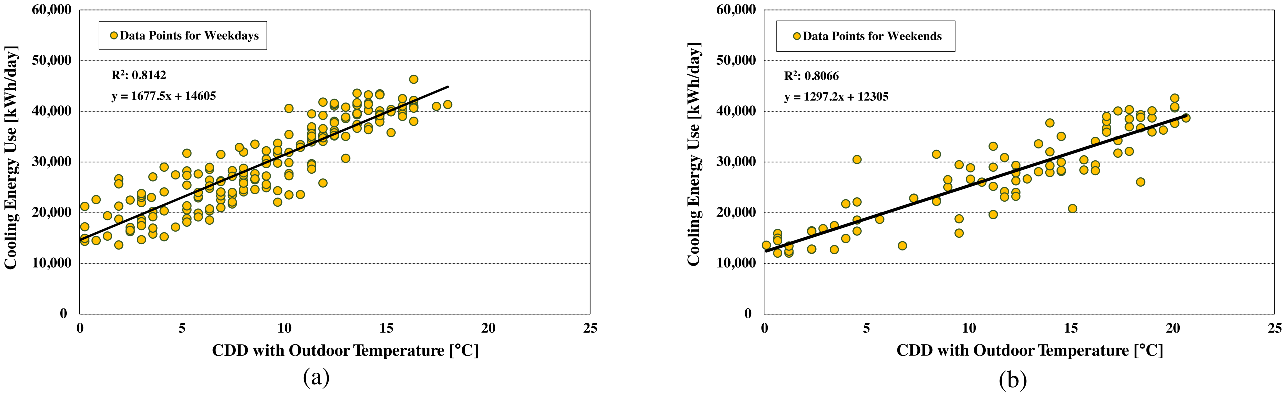


FIGURE 4. Cooling energy predictions with linear regression models and CDD [27]. The result of the correlation fits to the linear function shown in Eq. (1).

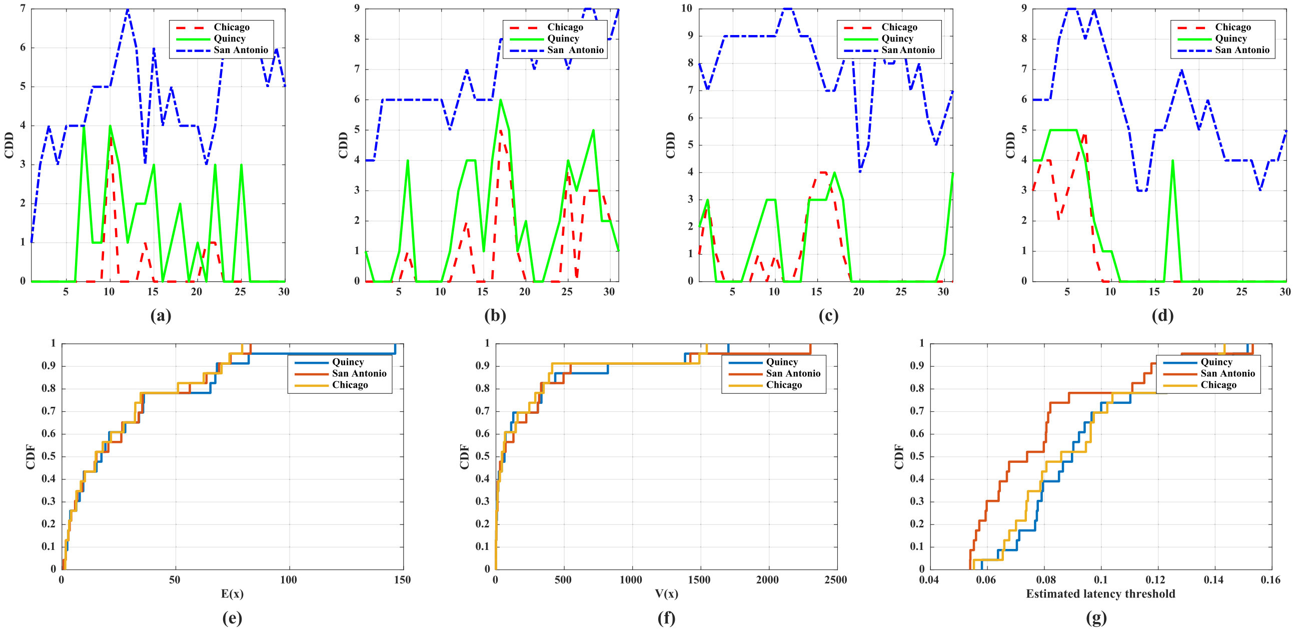


FIGURE 5. (a) to (d): The CDD of three Microsoft data center locations in four months: Apr, May, Jun, Jul. (e) and (f): Cumulative distribution of mean $E(x)$ and variance $V(x)$ for the requests' latency. (g): Cumulative distribution of latency threshold with $\mathrm{Pr}^{SLA} = 0.95$.

correlation: an equation and the R^2 value. Linear function (1) represents the correlation of cooling energy consumption and CDD. The R^2 is a measure of how good the correlation is. A good correlation between CDD and energy consumption indicates that the methodology is sound. In other words, the higher the R^2, the better.

Fig. 4 shows the cooling energy predictions with linear regression models and CDD [27]. Using Fig. 4(a) as an example, the linear function is $f(c) = 1677.5 * c + 14605$. The "$f(c)$" corresponds to the energy consumption. The "c" corresponds to the specific value of CDD. The parameter that multiplies the c represents the gradient of the trend line. The constant at the end is the intercept. In theory, this should represent the "baseload energy consumption". Most importantly the equation enables us to estimate cooling energy consumption from CDD. By plugging a known CDD into the equation, we can calculate the predicted cooling energy consumption for the time slot that the CDD covered. Fig. 5(a), (b), (c), (d) show the CDD of three data center locations in four months, respectively. These CDD data will serve as the parameter in VI-C for predicting energy consumption for reducing heat conducted with outside world.

C. EXPERIMENTAL RESULTS

We analyze the latency mean $E(x)$ and variance $V(x)$ of each request. Fig. 5(e) and (f) show the cumulative distribution of $E(x)$ and $V(x)$. The latency is based on millisecond, the average latency of more than 80% requests are less than 50 milliseconds and up to 90% requests have a better degree of aggregation. The value of μ and α can be obtained by solving functions (9) and (10). μ and α can be used for latency threshold estimation with Eq. (13). Fig. 5(g) shows the cumulative distribution of the estimated latency threshold with $p = 0.95$. In Fig. 5(g), the latency is based on second,

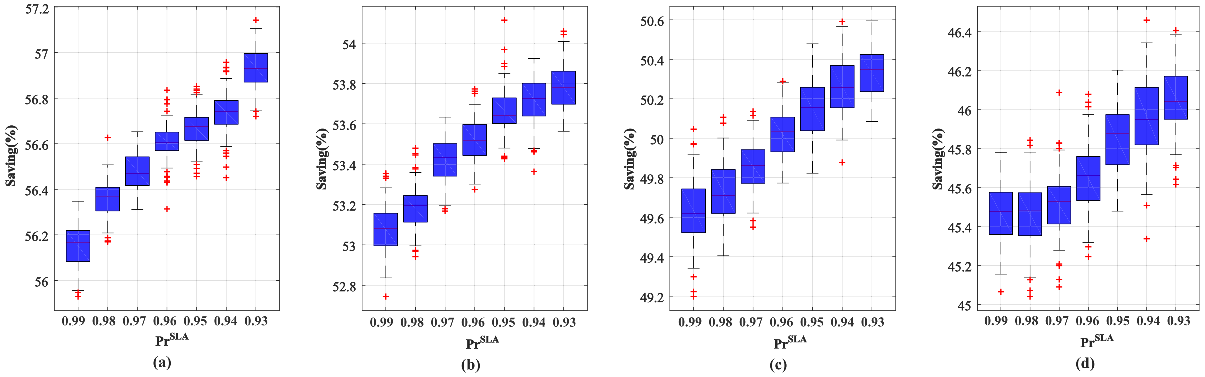


FIGURE 6. The total saving electric energy of three data centers in different correlation situations; (a) *a*=1677.5, *b*=14605; (b) *a*=1500, *b*=14000; (c) *a*=1297.2, *b*=12305; (d) *a*=1000, *b*=11000.

more than 95% requests are less than 0.15 second, which is consistent with Fig. 5(e). The three CDF graphs show that our theory for SLA latency estimation is feasible.

With SLA tail latency constraint and latency distribution of Cloud Services, we implement the prototype version of our design with dynamic programming, and continue the trace-based simulation by replaying the request of Cloud Services.

1) ENERGY CONSUMPTION

With the correlation of the cooling energy consumption and the CDD, we obtain the linear function $f(c)$. A set of a and b is used to simulate how our algorithm reschedules the workloads to target data center group. The situation of total saving electric energy for three data centers are shown in Fig. 6(a), (b), (c), (d). In each sub-figure, the saving is the summation of four months: Apr, May, Jun and Jul. The performance is related to the linear equation which means that the cooling energy consumption is in varying degrees with different correlation result. In fact, under workload rescheduling, the real-world data centers used in the simulation can save more than 40% cooling energy consumption.

2) TAIL LATENCY SLA REQUIREMENTS

One important feature of our design is that Cloud Service provides high-percentile tail latency. High-percentile tail latency enables our design to reduce the cooling energy consumption while Cloud Services can meet the tail latency requirements.

The value of $\Pr^{SLA}$ represents the tail latency satisfaction levels. If we set a higher value of $\Pr^{SLA}$, the customer will be satisfied, but the cooling energy consumption of the data center will increase because the choice of data center becomes narrow. If we set a lower value of $\Pr^{SLA}$, then the cooling energy consumption can be maintained a low-level, but QoS of Cloud Services will not be guaranteed which reflected in the decrease of the percentile that service latency does not exceed the latency threshold.

We changed the value of $\Pr^{SLA}$ from 0.93 to 0.99. The reduction is shown in Fig. 6(a), (b), (c), (d). As an example, Fig. 6(a) shows the relationship between tail latency percentile and saving electric energy. The saving is decreased if the Cloud Services have loosed tail latency constraints. If the Cloud Services have strict latency requirement, the Cloud Services can only be scheduled to a limited number of data centers. It means that if Cloud Services choose the data center group with least internal-outside temperature difference, we must realize the unexpected risk of not meeting the expected tail latency.

D. LATENCY PREDICTION ERROR ANALYSIS

Our design expects to schedule the Cloud Service to the lowest price data center as many as possible. However, since it does not take data center capacity and interference into consideration, a data center may become overloaded and hence may not meet the high-percentile tail latency constraint. The data center utilization can influence the performance of Cloud Services. If the data center utilization keeps at low, the error ratio of the probability tail latency estimation can tolerate. Otherwise, the error ratio will increase and as a result reduces the precision of workload rescheduling. It finally causes the added energy consumption. Fig. 6(a), (b), (c), (d) present different degrees of error caused by the latency estimation error. Hence, our design attempts to reschedule the workloads to data centers with least internal-outside temperature difference and with possible low utilization. If the data center is in high utilization, then the interference may enlarge the error ratio of the tail latency prediction.

VII. CONCLUSION

This work aims to reduce the cooling energy consumption of multiple data centers while guarantee the Cloud Service's SLA tail latency. The cooling energy of one data center is used to reduce the heat conducted with the outside world and the heat produced by the data center IT equipment. We correlate the historical energy bill and CDD, and work out the cooling energy consumption model by plugging in the energy utilization for reducing the heat generated by the IT equipment. We proposed a novel data-driven approach to determine the tail latency for different workload scheduling decisions, by profiling latency with G/G/m queue theorem at a low complexity. A weather-aware geo-scheduling algorithm is proposed, which proactively places workloads in data centers that with less temperature difference. The performance evaluation has been conducted with numerical studies and simulations. The result shows that our design can save more than 40% of the cooling energy consumption for multiple data centers while ensures the tail latency of Cloud Service meets the SLA constraint.

ACKNOWLEDGMENT

Jiahong Wu and Yuan Jin contributed equally to this work.

REFERENCES

[1] A. Shehabi *et al.*, "United states data center energy usage report," Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep. LBNL-1005775, 2016.

[2] Z. Song, X. Zhang, and C. Eriksson, "Data center energy and cost saving evaluation," *Energy Procedia*, vol. 75, pp. 1255–1260, Aug. 2015.

[3] BizEE Software. (2017). *Degree Days.net*. [Online]. Available: http://www.degreedays.net

[4] Microsoft Azure. (2017). *Azure Regions*. [Online]. Available: https://azure.microsoft.com/en-us/regions/

[5] D. Serrano *et al.*, "Towards QoS-oriented SLA guarantees for online cloud services," in *Proc. 13th IEEE/ACM Int. Symp. Cluster, Cloud, Grid Comput.*, May 2013, pp. 50–57.

[6] R. Sawyer, "Calculating total power requirements for data centers," White paper, American Power Conversion, vol. 562, 2004.

[7] A. Azimzadeh and N. Tabrizi, "A taxonomy and survey of green data centers," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2015, pp. 128–131.

[8] Y. Cui, C. Ingalz, T. Gao, and A. Heydari, "Total cost of ownership model for data center technology evaluation," in *Proc. 16th IEEE Int. Soc. Conf. Thermal Thermomech. Phenom. Electron. Syst. (ITherm)*, May/Jun. 2017, pp. 936–942.

[9] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 732–794, 1st Quart., 2016.

[10] J. Patrick. (2012). *Data Center Knowledge*. [Online]. Available: http://www.datacenterknowledge.com/archives/2012/10/16/where-will-the-next-silicon-valley-be/

[11] S. Sahana, R. Bose, and D. Sarddar, "Server utilization-based smart temperature monitoring system for cloud data center," in *Industry Interactive Innovations in Science, Engineering and Technology*. Singapore: Springer, 2018, pp. 309–319.

[12] T. Ding, Z. G. He, T. Hao, and Z. Li, "Application of separated heat pipe system in data center cooling," *Appl. Thermal Eng.*, vol. 109, pp. 207–216, Oct. 2016.

[13] R. Schmidt, M. Iyengar, and R. Chu, "Data centers: Meeting data center temperature requirements," *ASHRAE J.*, vol. 47, no. 4, pp. 44–46, 2005.

[14] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, "Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 11, pp. 1458–1472, Nov. 2008.

[15] M. A. Oxley *et al.*, "Rate-based thermal, power, and co-location aware resource management for heterogeneous data centers," *J. Parallel Distrib. Comput.*, vol. 112, pp. 126–139, Feb. 2018.

[16] E. Pakbaznia and M. Pedram, "Minimizing data center cooling and server power costs," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design (ISLPED)*, New York, NY, USA, 2009, pp. 145–150.

[17] T. Chen, X. Wang, and G. B. Giannakis, "Cooling-aware energy and workload management in data centers via stochastic optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 402–415, Mar. 2016.

[18] Z. Liu, M. Lin, A. Wierman, S. Low, and L. L. H. Andrew, "Greening geographical load balancing," *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 657–671, Apr. 2015.

[19] A. Rahman, X. Liu, and F. Kong, "A survey on geographic load balancing based data center power management in the smart grid environment," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 214–233, 1st Quart., 2014.

[20] S. Esfandiarpoor, A. Pahlavan, and M. Goudarzi, "Structure-aware online virtual machine consolidation for datacenter energy improvement in cloud computing," *Comput. Elect. Eng.*, vol. 42, pp. 74–89, Feb. 2015.

[21] Z. Cao and S. Dong, "Dynamic VM consolidation for energy-aware and SLA violation reduction in cloud computing," in *Proc. 13th Int. Conf. Parallel Distrib. Comput., Appl. Technol.*, Dec. 2012, pp. 363–369.

[22] Y. Guo, Y. Gong, Y. Fang, P. P. Khargonekar, and X. Geng, "Energy and network aware workload management for sustainable data centers with thermal storage," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 8, pp. 2030–2042, Aug. 2014.

[23] Z. Jiang, W. Huang, I. You, Z. Qian, and S. Lu, "Thermal-aware task placement with dynamic thermal model in an established datacenter," in *Proc. 8th Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput.*, Jul. 2014, pp. 1–8.

[24] F. Kaplan, J. Meng, and A. K. Coskun, "Optimizing communication and cooling costs in HPC data centers via intelligent job allocation," in *Proc. Int. Green Comput. Conf.*, Jun. 2013, pp. 1–10.

[25] O. Tuncer, K. Vaidyanathan, K. Gross, and A. K. Coskun, "CoolBudget: Data center power budgeting with workload and cooling asymmetry awareness," in *Proc. IEEE 32nd Int. Conf. Comput. Design (ICCD)*, Oct. 2014, pp. 497–500.

[26] V. Villebonnet and G. D. Costa, "Thermal-aware cloud middleware to reduce cooling needs," in *Proc. IEEE 23rd Int. WETICE Conf.*, Jun. 2014, pp. 115–120.

[27] M. Shin and S. L. Do, "Prediction of cooling energy use in buildings using an enthalpy-based cooling degree days method in a hot and humid climate," *Energy Buildings*, vol. 110, pp. 57–70, Jan. 2016.

[28] N. Rasmussen, "Calculating total cooling requirements for data centers," White paper, American Power Conversion, vol. 25, pp. 1–8, 2007.

[29] Wikipedia. (2017). *Coefficient of Performance (COP)*. [Online]. Available: https://en.wikipedia.org/wiki/Coefficient_of_performance

[30] Q. Pu *et al.*, "Low latency geo-distributed data analytics," in *Proc. ACM Conf. Special Interest Group Data Commun. (SIGCOMM)*, New York, NY, USA, 2015, pp. 421–434.

[31] A. Gupta *et al.*, "Mesa: Geo-replicated, near real-time, scalable data warehousing," *Proc. VLDB Endowment*, vol. 7, no. 12, pp. 1259–1270, Aug. 2014.

[32] A. Vulimiri, C. Curino, P. B. Godfrey, T. Jungblut, J. Padhye, and G. Varghese, "Global analytics in the face of bandwidth and regulatory constraints," in *Proc. 12th USENIX Conf. Netw. Syst. Design Implement. (NSDI)*, 2015, pp. 323–336.

[33] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha, "CSPAN: Cost-effective geo-replicated storage spanning multiple cloud services," in *Proc. ACM SIGCOMM Conf.*, 2013, pp. 545–546.

[34] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for Internet-scale systems," in *Proc. ACM SIGCOMM Conf. Data Commun.*, New York, NY, USA, 2009, pp. 123–134.

[35] 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*, Mar. 2010, pp. 1–9.

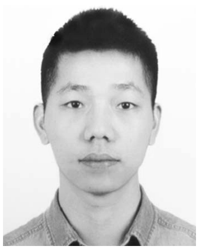
[36] R. D. Gupta and D. Kundu, "Theory & methods: Generalized exponential distributions," *Austral. New Zealand J. Stat.*, vol. 41, no. 2, pp. 173–188, 1999.

[37] M. Nguyen, Z. Li, F. Duan, H. Che, Y. Lei, and H. Jiang, "The tail at scale: How to predict it?" in *Proc. 8th USENIX Conf. Hot Topics Cloud Comput. (HotCloud)*, Berkeley, CA, USA, 2016, pp. 120–125. [Online]. Available: http://dl.acm.org/citation.cfm?id=3027041.3027061

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[39] Microsoft. (2017). *Microsoft Datacenters*. [Online]. Available: http://www.microsoft.com/en-us/server-cloud/cloud-os/global-datacenters.aspx

[40] G. Chen *et al.*, "Energy-aware server provisioning and load dispatching for connection-intensive Internet services," in *Proc. 5th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Berkeley, CA, USA, 2008, pp. 337–350.



JIAHONG WU received the B.E degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. He is currently a Graduate Student with the School of Software, Shanghai Jiao Tong University, China. His research interests mainly include cloud computing, real-time system, and networking.

YUAN JIN received the M.S. degree from McGill University, Montreal, QC, Canada, in 2011. He was a Visiting Research Scholar with the School of Software, Shanghai Jiao Tong University, China. His research interests mainly include distributed computing, real-time systems, and cyber-physical systems.

JIANGUO YAO received the B.E., M.E., and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2004, 2007, and 2010, respectively. He was a Joint Post-Doctoral Fellow with the Ecole Polytechnique de Montreal and McGill University in 2011. He is currently with Shanghai Jiao Tong University as an Associate Professor. His research interests are cloud computing, industrial big data, and cyber-physical systems.