Minimizing Electricity Cost and Water Footprint for Geo-Distributed Interactive Services with Tail Latency Constraint

Executive Summary

In the last two decades, data centers (DC) have been expanding significantly in both number and scale by service providers such as Microsoft, Google, Facebook, and Amazon. Overall US data centers consumed about 2% of the total U.S. electricity consumption and 166 billion gallons of water in 2014, which includes both water consumed directly at data center sites and water used to generate the electricity that powered them. DCs are major energy consumers among all industry sectors, constituting a large fraction of their operators' expenses. On the other hand, many data centers are in “high” and “very high” water stress regions and place significant water impact on local sustainability.

Due to costly cross-border data movement and data sovereignty, data replication is available only in data centers within a region that practically is a large country or continent. A soaring amount of interactive services (e.g., real-time global user/market analysis) must be processed at multiple data center locations simultaneously. There are differences in electricity prices and water efficiencies in difference data center locations. Moving optimal workloads to cheap data center locations can reduce the total cost of electricity and water footprint. If too much workload is queuing at cheap data center locations, queuing time will make the response time too long to be accepted by end users. Besides, both service requests' arrivals and demand times are randomly distributed. In this paper, an optimization is designed to minimize the total cost of electricity and water footprint without sacrificing users' usability. Real workload data from Google is used to conduct optimization simulation, and it is shown that about 10% total reduction of electricity and water footprint can be achieved when compared to performance-aware but cost-oblivious approach. My work advances the prior research by adding water footprint as a new metric.

Minimizing Electricity Cost and Water Footprint for Geo-Distributed Interactive Services with Tail Latency Constraint

Abstract

An early effort is made to minimize the total cost of electricity and water footprint in data centers for geo-distributed interactive services, which rely on request processing in multiple data centers due to distributed data sets and are subject to a tail latency constraint. It extends the prior research by adding water footprint as a new metric. A parameterized total cost of electricity and water footprint is formulated with a weighing parameter. A probability-based latency threshold satisfaction is used for the tail latency constraint. A geographic load balancing technique is used to exploit spatial and temporal variations in electricity prices and water efficiencies in different data center locations. Using a data-driven approach at runtime, the tail latency probability is obtained by profiling network latency and data center latency statistics and convolution of the two probability mass functions, and a MATLAB programming solver is used to make optimal dispatch decisions of random service requests. Using real workload data from Google, a traced-based discrete-event simulation is conducted to validate performance, showing the total cost reduction of electricity and water footprint can be achieved by 10.04% on average when compared to performance-aware but cost-oblivious approach. The impact of weighing parameter of water footprint and response latency tail percentile is evaluated too.

Minimizing Electricity Cost and Water Footprint for Geo-Distributed Interactive Services with Tail Latency Constraint

I. Introduction

The world is experiencing a dramatic socio-technical change. In the last two decades, data centers (DC) have been expanding significantly in both number and scale by service providers such as AT&T, Microsoft, Google, Facebook, and Amazon [1], [2]. Running a large, industrial-scale data center needs a huge amount of electricity and water. Overall US data centers consumed about 2% of the total U.S. electricity consumption [3], and 166 billion gallons of water in 2014 [4], which includes both water consumed directly at data center sites and water used to generate the electricity that powered them. Although growth in data center energy consumption and water footprint in the U.S. has slowed down since 2007, DCs are still major energy consumers among all industry sectors, constituting a large fraction of their operators' expenses [5], [6], [7], [8], [9]. On the other hand, many data centers are in "high" and "very high" water stress regions [10] and place significant water impact on local sustainability. This situation drew a lot of public attention, as the drought in California grew especially acute in the summer of 2015. Nevertheless, hyper-scale data centers built by Internet and cloud giants continue to grow [3], [4]. This fact motivates the research in this paper.

While service providers have made many engineering optimizations by using highly efficient equipment, researchers have made tremendous progress in optimizing energy consumption and water footprint of data centers through software optimization (i.e., workload balancing algorithms) [11]-[14]. In geo-distributed data centers [12], spatial variations across data centers worldwide and temporal variations within each data center have been exploited under random workload sequences. These spatial variations include different electricity prices, available renewable energies, carbon efficiencies, water prices, water usage efficiencies, and others in different geographical locations. The temporal variations include hourly and seasonal variations of humidity, temperature, and the values just mentioned in spatial variations [1]. Different types of workloads in DCs have been studied, including batch online requests (e.g., large bank transactions) [14] and interactive services (e.g., Google searches, global market analysis, etc.) [12]. Centralized and geo-distributed processing of interactive services has been studied in the situations of a centralized data center [8], [15], [16], [19], [20] and distributed data sets [12], [21], [22], respectively. An important aspect of data center management is that end-to-end response time (i.e., total latency) to a client's interactive service request should be less a threshold, as a performance or service layer agreement (SLA) constraint. Two different total latency constraints are used in interactive service studies: a tail latency constraint (e.g., p95 latency, i.e., at least 95% of the requests should have a latency not exceeding a certain threshold) has been used as SLA to ensure consistently low latencies [12], [23], and an average latency constraint is that the average total response time should be less than a certain threshold [8], [15], [16], [19], [20]. All the variations mentioned above are exploited in so-called geographic load balancing techniques (GLB) by many prior studies for various design purposes, such as reducing electricity cost, maximizing the utilization of renewable energy, and reducing carbon analysis [6]-[8], [15]-[18]. With all these promising progress, however, very little has been published on total cost reduction of both energy [12] and water footprint [11] in geo-distributed interactive services, with consideration of number of active servers in data centers [11], and this research aims at filling this knowledge gap.

This paper extends [12] by adding water footprint as a new metric, i.e., minimizing the total cost of both electricity and water footprint in all data centers subject to a tail latency constraint is researched. An interactive service that relies on geo-distributed data sets [12] is considered. Each service request is sent to all regions simultaneously, and only one data center in each region is selected to process the request. An overview of a typical geo-distributed interactive service is illustrated in Fig. 1. Due to costly cross-border data movement and data sovereignty, data replication is available only in data centers within a region that practically is a large country or continent [12].

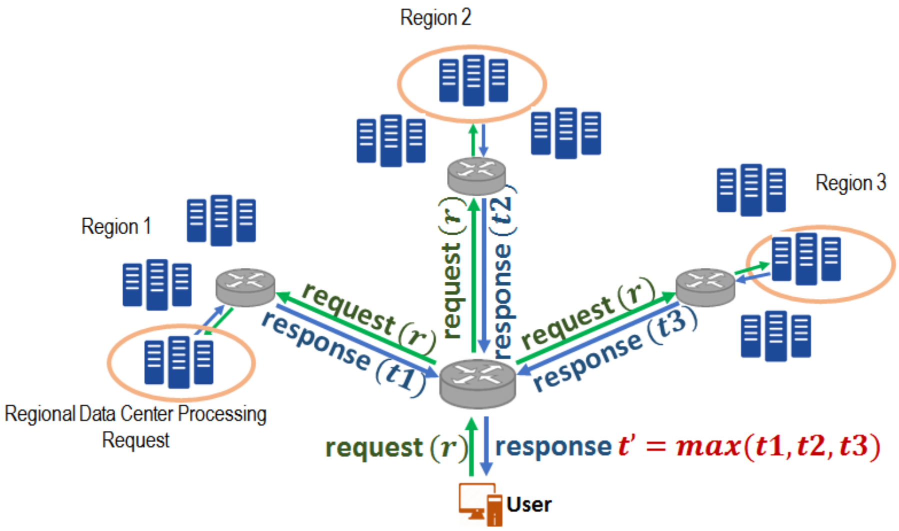


Fig. 1. Geo-distributed interactive service. The response time is equal to the longest response time [12].

Data centers are considered homogeneous, with the only difference being the number of active servers that are turned on. A discrete event simulator based on DESMO-J [25] is implemented to conduct a trace-based simulation of the interactive services' network traffic and data center processing, and a MATLAB program is developed to make dispatch decisions of random service requests. Finally, real data from Google is used to conduct performance optimization evaluation, and it is shown that about 10% total reduction of electricity and water footprint can be achieved.

II. Problem Formulation

An interactive service provider operates N data centers around the world, represented by $j = \{1, 2, \cdots, N\}$, where there are m_j servers turned on in data center j. There are M different geographical regions with N_m data centers in region m, where $m = \{1, 2, \cdots, M\}$. The total number of data centers is $\sum_{m=1}^{M} N_m = N$. There are S different traffic sources from which users make service requests. A time-slotted model is considered, where a decision is made every time slot (e.g., 15 minutes) and the workload arrival is assumed known for the next time slot.

A. Electricity Consumption and Cost in Data Centers

Mathematically, the total server electricity consumption of data center j at time t can be expressed as a linear function of its total workload and number of servers turned on [11], [12]. That is,

$$e_j\big(a_j(t), m_j(t)\big) = m_j(t)\left[e_{0,j} + e_{c,j}\frac{a_j(t)}{m_j(t)\mu}\right] \tag{1}$$

where $a_j(t)$ is the total workload dispatched to data center j, $m_j(t)$ is the number of servers that are turned on at time t, $e_{0,j}$ is static power consumption by a server in data center j, $e_{c,j}$ is computing power consumption by a server in data center j, and μ is one server's capacity. The incurred electricity cost of data center j can be expressed as [11]

$$q_j(t)\left[\gamma_j(t) \cdot e_j\left(a_j(t), m_j(t)\right) - R_j(t)\right]^+ \tag{2}$$

where $[.]^+ = \max\{., 0\}$, $R_j(t)$ is on-site renewable energy, $\Upsilon_j(t)$ is PUE of data center j at time t, and $q_j(t)$ is electricity price at data center j location and at time t. When $R_j(t)$ is greater than the total energy needed by data center j, electricity cost is zero. Power usage effectiveness is defined as

$$PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy} = 1 + \frac{Non\ IT\ Equipment\ Energy}{IT\ Equipment\ Energy} \quad (3)$$

B. Water Footprint at Data Centers

The water footprint in this paper focuses on direct water usage in DC cooling technology and indirect water usage in electricity generation. As industry standards, direct water usage effectiveness is defined as [1], [11]

$$WUE = \frac{Direct\ Water\ Usage}{IT\ Equipment\ Energy} \quad (4)$$

and energy water intensity factor is defined as [11]

$$EWIF = \frac{Indirect\ Water\ Usage}{ITEquipment\ Energy - Renewable} \quad (5)$$

The water footprint of data center j at time t can be expressed as [11]

$$w_j(t) = \varepsilon_{j,d}(t) \cdot e_j\left(a_j(t), m_j(t)\right) + \varepsilon_{j,id}(t) \cdot \left[\gamma_j(t) \cdot e_j\left(a_j(t), m_j(t)\right) - R_j(t)\right]^+ \quad (6)$$

where $\varepsilon_{j,d}(t)$ is the direct WUE of data center j at time t and $\varepsilon_{j,id}(t)$ is EWIF of data center j.

C. Total Parameterized Cost

A parameterized total cost function of data center j at time t, can be expressed as [14]

$$f_j(t) = q_j\left[\gamma_j(t) \cdot e_j\left(a_j(t), m_j(t)\right) - R_j(t)\right]^+ + h_w w_j(t) \quad (7)$$

where $h_w \geq 0$ is the weighing parameter for water footprint. This multi-objective formulation is common in the literature [14]. Hence, the total cost over the simulation duration can be integrated as

$$Total\ Cost = \sum_{t=1}^{T}\left(\sum_{j=1}^{N} f_j(t)\right) \tag{8}$$

where T = number of time slot during simulation time period.

D. Latency Constraint and Profiling

The tail latency performance of source i is expressed as $p_i(\vec{a}, \vec{r_i})$, a function of data center workload $\vec{a} = \{a_1, a_2, \dots, a_N\}$ and network route/path $\vec{r_i} = \{r_{i,1}, r_{i,2}, \dots, r_{i,N}\}$, where $r_{i,j}$ denotes the network route from source i to data center j. It is noted that p_i represents the probability Pr $(d_i \leq D_i)$ that the end-to-end response time d_i for requests from source i does not exceed the threshold value D_i (e.g., 150ms). The tail latency constraint p95 is expressed as

$$p_i(\vec{a}, \vec{r_i}) \geq P_i^{SLA}, \forall i = 1, \cdots, S \tag{9}$$

Each service request is simultaneously sent to all the M geographical regions, and only one data center from each region is selected for processing the request. This yields $G = \prod_{m=1}^{M} N_m$ possible data center groups for a request. At the source i, a load distribution decision vector is defined as

$$\vec{\lambda_i} = \{\lambda_{i,1}, \lambda_{i,2}, \dots, \lambda_{i,G}\}^T \tag{10}$$

where $\lambda_{i,g} \geq 0$ denotes the amount of requests sent to group g from source i. The total workload from source i is

$$\Lambda_i = \sum_{g=1}^{G} \lambda_{i,g} \tag{11}$$

Considering all the traffic sources, the workload distribution matrix is defined as

$$\lambda = \{\vec{\lambda_1}, \vec{\lambda_2}, \dots, \vec{\lambda_S}\} \tag{12}$$

Hence, the total workload sent to data center j is

$$a_j = \sum_{i=1}^{S} \sum_{g \epsilon \mathscr{g}_j} \lambda_{i,g} \tag{13}$$

where $\mathcal{g}_j$ represents the set of data center groups that have data center j as an element. Since there are S sources and N data centers, there are a total of $R = S \times N$ routes, each representing a network path from a source location to a data center location. Based on the latency independence property [12] (i.e., an end-to-end response time of requests sent along one route is practically independent of that along another route), it is possible that the response time probabilities along different routes can be combined to give the response time probability for requests from source i to each data center group. That is, [12]

$$p_{i,g}^{group}(\vec{a}, \vec{r_i}) = \prod_{j \epsilon \mathcal{J}} p_{i,j}^{route}(a_j, r_{i,j}, m_j) \tag{14}$$

where $\mathcal{J}$ is the set of data centers that are in the data center group g. Hence the probability of $(d_i < D_i)$ at source i can be the weighted average across all the involved data center groups as [12]

$$p_i(\lambda) = P_i\ (\vec{a}, \vec{r_i}) = \frac{1}{\Lambda_i} \sum_{g=1}^{G} \lambda_{i,g}\ P_{i,g}^{group}(\vec{a}, \vec{r_i}) \tag{15}$$

Similar to [12], the end-to-end latency distribution for $r_{i,j}$ can be obtained by

$$F_{i,j}^{R} = F_{i,j}^{N} *\ F_j^D(x) \tag{16}$$

where $F_{i,j}^{N}$ is the network latency distribution of route $r_{i,j}$, $F_j^D(x)$ is the data center latency distribution with load x, and $" * "$ is the convolution operator. This makes it easy to calculate $p_{i,j}^{route}(a_j, r_{i,j})$, and hence $p_{i,g}^{group}(\vec{a}, \vec{r_i})$ and $P_i\ (\vec{a}, \vec{r_i})$ in (14) and (15). A total of $S \times N$ network latency distributions and $N \times W$ data center latency distributions need to be profiled, where W represents the levels of workload for each route. It is noted that the load distribution matrix $\lambda = \{\overrightarrow{\lambda_1}, \overrightarrow{\lambda_2}, \dots, \overrightarrow{\lambda_S}\}$ is the main decision variable in this problem.

E. Updated Problem Formulation

In summary, the problem is formulated as

GLB-2-WF:

$$minimize_{(\lambda,m_j,h_w)} \sum_{j=1}^{N} \left(q_j \left[\gamma_j(t) \cdot e_j \left(a_j(t), m_j(t) \right) - R_j(t) \right]^{+} + h_w w_j(t) \right) \quad (17)$$

$$\text{Subject to} \quad p_i(\lambda) \geq P_i^{SLA} \text{ for source } i \quad (18)$$

$$\sum_{g=1}^{G} \lambda_{i,g} = \Lambda_i, \ \forall i = 1, 2, \dots, S \quad (19)$$

$$a_j \leq m_j \mu, \forall j \in N \quad (20)$$

where the constraint (18) is a nonlinear part of this optimization problem, and (19) ensures that all requests from a traffic source are processed.

To solve the problem GLB-2-WF, as shown in Fig. 2, an algorithm similar to McTail in [12] is used to periodically determine the tail latency probability and meet the constraints (18), while exploiting electricity cost and water footprint diversities in different data centers to minimize the total cost (17). The input to McTail includes the profiled network latency probability mass function and data center latency probability mass function, the estimated workload arrival at each source at time t, and diversified parameters in each data center location. A program solver in MATLAB can be used to solve the problem GLB-2-WF, and outputs the optimized GLB decisions that optimally split the incoming workloads at each source to different geo-distributed data centers for processing. The response time statistics of each data center is profiled by a data-driven approach in a M/M/1 discrete event simulation. The network latency statistics are in half-normal distributions, where the mean and standard deviation depend on the distance between the source and the data center. Then, the probability of $(d_i < D_i)$ at source i can be easily calculated from (16), (14) and (15). The response time distributions need to be updated in case of significant latency change (network latency or data center latency) [12].

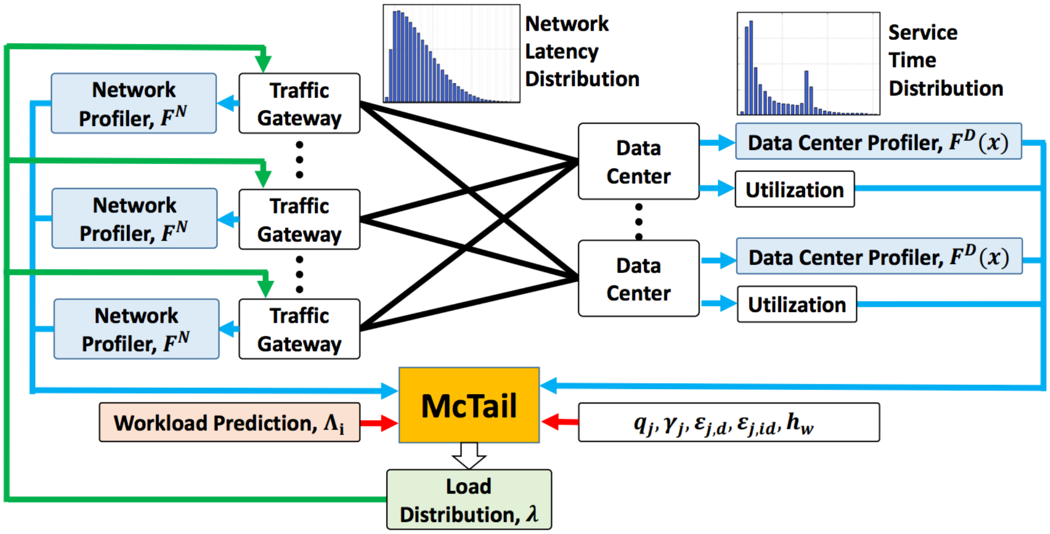

Network Latency Distribution
Service Time Distribution
Network Profiler, F^N
Traffic Gateway
Network Profiler, F^N
Traffic Gateway
Network Profiler, F^N
Traffic Gateway
Data Center
Data Center Profiler, $F^D(x)$
Utilization
Data Center
Data Center Profiler, $F^D(x)$
Utilization
McTail
Workload Prediction, Λ_i
$q_j, \gamma_j, \varepsilon_{j,d}, \varepsilon_{j,id}, h_w$
Load Distribution, λ

Fig. 2. McTail: Network latency probability mass function (PMF) and data center latency PMF are profiled separately and sent to McTail periodically [12].

III. Performance Evaluation

Trace-based simulations on a DESMO-J discrete event simulator [25] are conducted for profiling the response time statistics of each data center, where there is one source and one data center (i.e., M/M/1 discrete event simulation). For validating SLA and calculate total cost over simulation duration of 24 hours, there are multiple sources and multiple data centers in multiple regions. The GLB-2-WF optimized decision is determined in MATLAB every hour. The simulator takes as inputs the service times and network latency distributions, which it then uses to simulate queuing and request processing.

The performance evaluation settings include two traffic sources: one in Chile, South America and one in Sydney, Australia, two DC regions: North America (USA) and Asia (China), and two data centers in each region: Oregon and North Carolina in North America and Beijing and Hong Kong in Asia.

To facilitate the programming in MATLAB and DESMO-J simulator, the equations in GLB-2-WF are rewritten in matrix format. First, a configuration matrix C, representing the relation of the data center groups and data centers, the workload distribution matrix λ (12), representing the main decision variables, and a functional matrix I can be expressed, respectively, as

$$C = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \qquad \lambda = \begin{pmatrix} \lambda_{1,1} & \lambda_{2,1} \\ \lambda_{1,2} & \lambda_{2,2} \\ \lambda_{1,3} & \lambda_{2,3} \\ \lambda_{1,4} & \lambda_{2,4} \end{pmatrix}, \quad I = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{21}$$

where the value $c_{j,g}$ is 1 if the data center group g has data center j, or 0 if the data center group g does not have data center j, and the value $\lambda_{i,g}$ is the amount of requests sent to group g from source i. Then, as shown in (22) from the left to the right, the configuration matrix multiplies the workload distribution matrix λ, yielding a matrix whose row values can be summed up across all traffic sources (by multiplying the functional matrix I) to give the total workload at a data center. Thus the total workload at each data center can be expressed by the main decision variable $\lambda_{i,j}$, which is nothing but (13).

$$C \cdot \lambda \cdot I = \begin{pmatrix} \lambda_{1,1} + \lambda_{1,2} + \lambda_{2,1} + \lambda_{2,2} \\ \lambda_{1,3} + \lambda_{1,4} + \lambda_{2,3} + \lambda_{2,4} \\ \lambda_{1,1} + \lambda_{1,3} + \lambda_{2,1} + \lambda_{2,3} \\ \lambda_{1,2} + \lambda_{1,4} + \lambda_{2,2} + \lambda_{2,4} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} \tag{22}$$

From (22), GLB-2-WF (17)-(20) can then be written as a function of the main decision variables $\lambda_{i,j}$ as follows.

$$Minimize_{\lambda,m_j}\left[(X_1+X_2+X_3+X_4)+(Y_1\quad Y_2\quad Y_3\quad Y_4)\cdot\begin{pmatrix}\lambda_{1,1}+\lambda_{1,2}+\lambda_{2,1}+\lambda_{2,2}\\ \lambda_{1,3}+\lambda_{1,4}+\lambda_{2,3}+\lambda_{2,4}\\ \lambda_{1,1}+\lambda_{1,3}+\lambda_{2,1}+\lambda_{2,3}\\ \lambda_{1,2}+\lambda_{1,4}+\lambda_{2,2}+\lambda_{2,4}\end{pmatrix}\right] \quad (23)$$

where,

$$\left.\begin{aligned}X_j &= m_j e_{0,j}\left[q_j\gamma_j+\left(\varepsilon_{j,d}+\varepsilon_{j,id}\gamma_j\right)h_w\right]-\left(q_j+\varepsilon_{j,id}\right)R_j\\ Y_j &= \frac{e_{c,j}}{\mu}\left[q_j\gamma_j+\left(\varepsilon_{j,d}+\varepsilon_{j,id}\gamma_j\right)h_w\right]\end{aligned}\right\} when\ \gamma_j\left(m_j e_{0,j}+\frac{e_{c,j}}{\mu}a_j\right)-R_j>0 \quad (24)$$

$$\left.\begin{aligned}X_j &= m_j e_{0,j}\varepsilon_{j,d}h_w\\ Y_j &= \frac{e_{c,j}}{\mu}\varepsilon_{j,d}h_w\end{aligned}\right\} when\ \gamma_j\left(m_j e_{0,j}+\frac{e_{c,j}}{\mu}a_j\right)-R_j\le 0 \quad (25)$$

subject to $P_i=\frac{1}{\Lambda_i}\left[\lambda_{i,1}P_{i,1}^{Group}+\lambda_{i,2}P_{i,2}^{Group}+\lambda_{i,3}P_{i,3}^{Group}+\lambda_{i,4}P_{i,4}^{Group}\right]\ge P_i^{SLA}$ for source i (26)

$$P_{i,1}^{Group}=P_{i,r_{i,1}}^{Route}(a_1,m_1)\cdot P_{i,r_{i,3}}^{Route}(a_3,m_3) \quad (27)$$

$$P_{i,2}^{Group}=P_{i,r_{i,1}}^{Route}(a_1,m_1)\cdot P_{i,r_{i,4}}^{Route}(a_4,m_4) \quad (28)$$

$$P_{i,3}^{Group}=P_{i,r_{i,2}}^{Route}(a_2,m_2)\cdot P_{i,r_{i,3}}^{Route}(a_3,m_3) \quad (29)$$

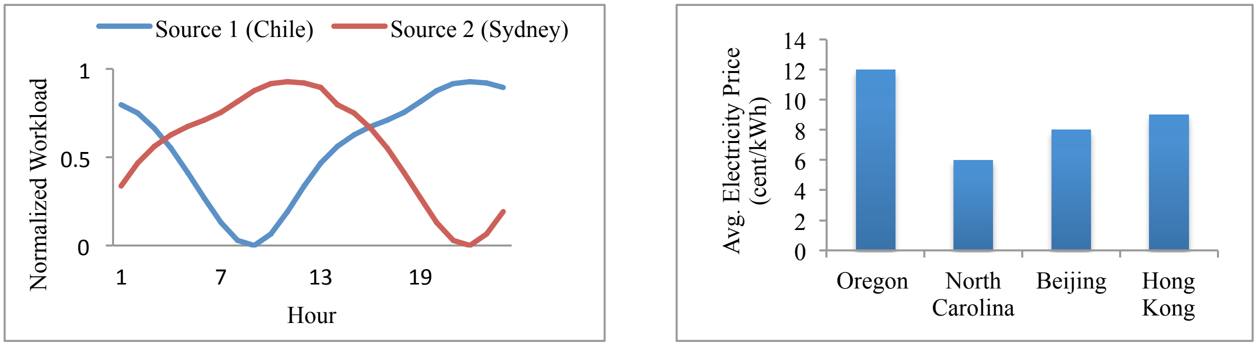
$$P_{i,4}^{Group}=P_{i,r_{i,2}}^{Route}(a_2,m_2)\cdot P_{i,r_{i,4}}^{Route}(a_4,m_4) \quad (30)$$

$$\Lambda_i=\lambda_{i,1}+\lambda_{i,2}+\lambda_{i,3}+\lambda_{i,4},\ \forall i=1,2 \quad (31)$$

$$\begin{pmatrix}\lambda_{1,1}+\lambda_{1,2}+\lambda_{2,1}+\lambda_{2,2}\\ \lambda_{1,3}+\lambda_{1,4}+\lambda_{2,3}+\lambda_{2,4}\\ \lambda_{1,1}+\lambda_{1,3}+\lambda_{2,1}+\lambda_{2,3}\\ \lambda_{1,2}+\lambda_{1,4}+\lambda_{2,2}+\lambda_{2,4}\end{pmatrix}\le\begin{pmatrix}m_1\mu\\ m_2\mu\\ m_3\mu\\ m_4\mu\end{pmatrix} \quad (32)$$

The source workload traces are taken from Google's Gmail service [24], and the trace data specifies the average normalized arrival rate over time from each source [12]. The workloads are scaled so that all the data centers operate at around 30% capacity on average.

The SLA threshold for p95 response time is chosen to be 1.25 seconds (i.e., at least 95% of requests have response time less than 1.25 seconds). For network latencies, half-normal distributions are used, with the mean dependent on the distance between the source and the data center and the standard deviation set to one. Approximately, a network latency of 1.64 milliseconds per 200 miles is used to calculate the mean, mirroring the real world [12]. This gives the mean a range of 68 to 187 milliseconds. The service demand follows an exponential distribution with the mean being 200 milliseconds.

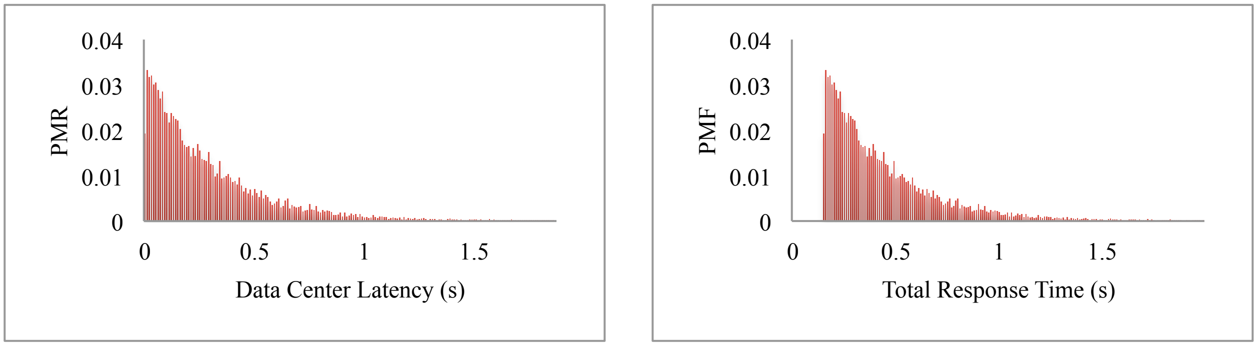


(a) (b)

Fig. 3. (a) The workload traces from Google's data for the two traffic sources, (b) Electricity prices at data center locations.

Workload traces and electricity prices are shown in Fig. 3. Electricity prices are taken from real world data at the locations of the data centers. Water and power efficiency data (i.e., WUE, PUE and EWIF) at different data center locations are taken from [27]-[36]. The weighing parameter h_w for water footprint is set at 0.05 in this study, which makes the electricity cost and water footprint cost remain approximately equal on average so that both electricity and water footprint are equally considered during the optimization.

The ratio of the static power consumption versus the dynamic power consumption of a server computer is 1:20 [26]. McTail optimization does not impact the static energy, even though it is very small compared to the dynamic energy consumption. Renewable energy is set to zero.

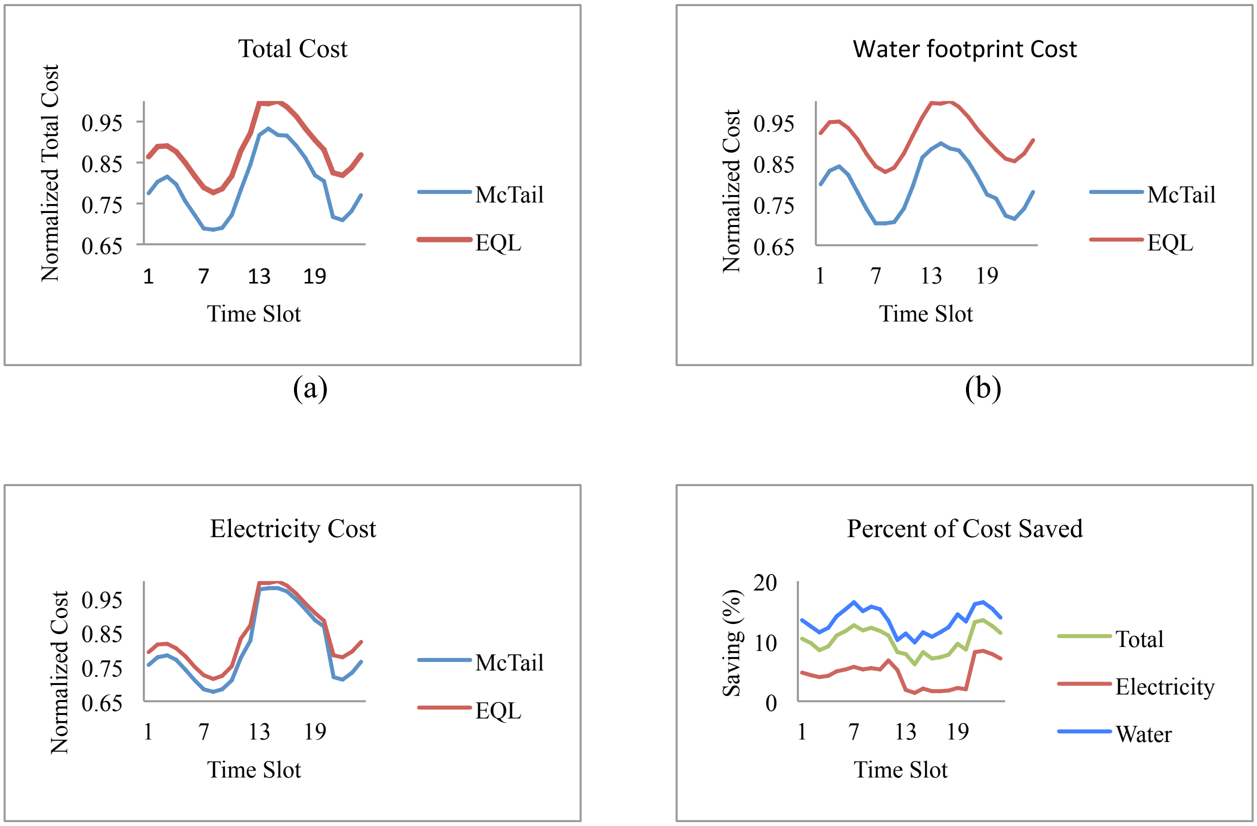


(a) (b)

Fig. 4. Data center latency distribution (a) and end-to-end response time distribution (b) under 30% data center utilization.

Fig. 4 shows the data center latency distribution (queuing delay plus service time) and the end-to-end Sydney-Oregon response time distribution that is obtained through a convolution of data center latency distribution and network latency distribution. The performance of McTail is compared to that of EQL (EQual Load distribution). EQL distributes the workloads equally among all data center groups.

Fig. 5 shows the normalized cost results. Fig. 5(a)-(c) show that McTail has lower costs than EQL through the simulation period since it exploits the difference in electricity prices and water efficiencies in different data center locations and balances workloads across DCs to achieve lower cost than EQL. Fig. 5(d) shows that, when compared to that of EQL, the total cost (electricity + water footprint) saved is 10.04%, the electricity costs saved is 4.49%, and the water footprint cost saved is 13.42% on average. By balancing workloads to low cost data centers, the overall cost can be reduced. With more data centers in the system under consideration, it is expected to achieve better savings, if the additional data centers have more varying diversities. It is noted that the savings for the water footprint are consistently greater than those of the electricity savings. This is due to increased diversities in water efficiencies across the data centers, while electricity costs are much closer in a single region (e.g. the WUE in Beijing differs from the WUE in Hong Kong by approximately 30%, but the electricity prices in the two regions only differ by around 10%). The SLA constraint is verified by using DESMO-J simulation. Fig. 6 shows the probabilities of response time less than 1.25s in each data center over 24 hours, and they are always greater than 95%. It is noted that the probability of latency threshold satisfaction in one source Sydney location is very close to 95% at all times. This is straightforward because otherwise McTail can achieve higher cost savings by moving additional workloads to cheaper data center locations as long as the probability of delay threshold satisfaction for all sources remains above 95%.



(a)

(b)

(c)

(d)

Fig. 5. Comparison of cost between McTail and EQL. (a) Normalized total cost. (b) Normalized electricity cost. (c) Normalized water footprint cost. (d) Percentage of cost saving.

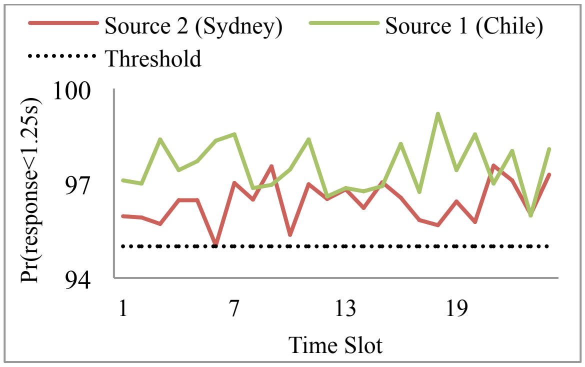


Fig. 6. Probabilities of end-to-end response time less than the 1.25s SLA threshold at traffic sources.

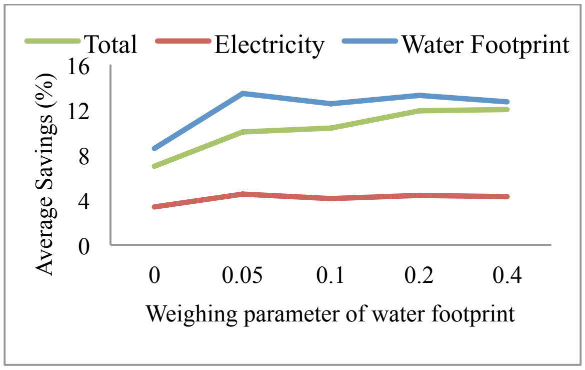
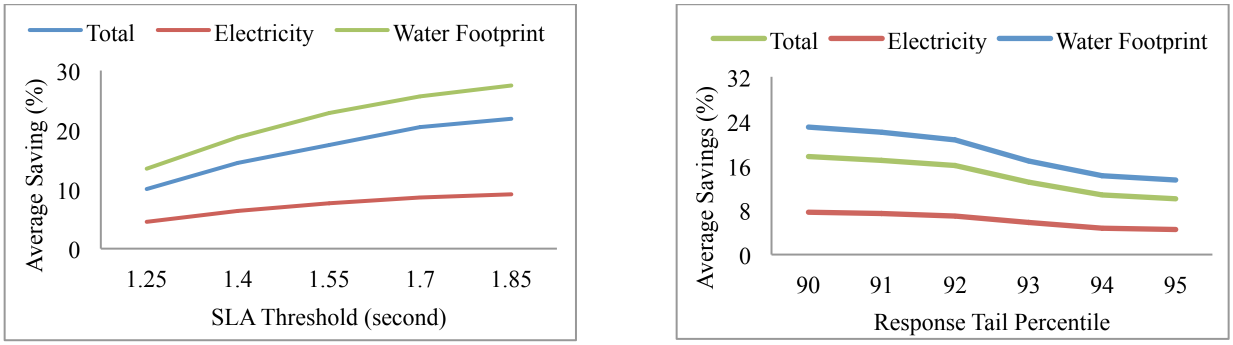


Fig. 7. Impact of weighing parameter of water footprint h_w

In Fig. 5, h_w equals 0.05 in order to equally consider both water footprint and electricity costs. Fig. 7 shows the impact of weighing parameter of water footprint on average saving percentage. If h_w is increased, simulation results show that the average total cost saving percentage is increased, and vice versa. The average electricity saving percentage remains almost constant with varying h_w, and average water footprint saving percentage increases slightly with increasing h_w. This effect can be explained as that an increased weighing parameter h_w means a higher priority in water footprint in the optimization algorithm, and that water efficiencies vary more among locations than electricity cost as stated above. The peak in the water footprint saving percentage at $h_w = 0.05$ is understood as a result of MATLAB optimization.



(a) (b)

Fig. 8. Impact of (a) SLA threshold and (b) tail percentile SLA target P_i^{SLA}

Fig. 8(a) shows the average total cost saving percentage at different SLA response time thresholds from 1.25 to 1.85 seconds and at the tail percentile 95. It is seen that when the response time threshold is increased, better cost saving is achieved. This effect is straightforward as with a relaxed threshold, more service requests can be dispatched to cheap data center locations without violating the SLA. Fig. 8(b) shows the impact of response tail percentage SLA target p_i^{SLA} on average saving percentage. The response time threshold is fixed at 1.25 seconds, and the tail percentile P_i^{SLA} varies from 90 to 95. As expected, a relaxed constraint results in increased savings as McTail has more flexibility in balancing workloads to low-cost DC locations.

IV. Future Work

In this paper, renewable energy is chosen to be zero, and the number of active servers that are turned on is chosen to be the same in all data centers. In the very next step, more simulation experiments will be conducted to research the impact of renewable energy and different number of servers in different data centers on the total cost of electricity and water footprint in geo-distributed interactive service data centers. In addition, this paper provides an exciting opportunity for service providers to exploit energy saving and water sustainability in their data center systems.

V. Conclusion

My work extends the prior research [12] by adding water footprint as a new metric. A preliminary and early effort is made to minimize the total cost of electricity and water footprint in data centers for geo-distributed interactive service subject to a tail latency constraint, by solving GLB-2-WF decisions. After the problem was formulated for geo-distributed interactive services which rely on request processing in multiple data centers due to distributed data sets, GLB-2-WF was solved by a programming solver in MATLAB and an event-based simulation was conducted to validate performance, showing it can reduce the total cost of electricity and water footprint by 10.04% on average when compared to performance-aware but cost-oblivious approach.

VI. Acknowledgement

I would like to thank my mentors for introducing me into this interesting and challenging problem of interactive services in geo-distributed data centers. The mathematical problems of stochastic workload, probability distributions, discrete event simulation, multivariate optimization with constraints, and many others, make this project a venture, yet the most enjoyable and fruitful learning journey in my life. This project has truly inspired me to pursue a science career and has shown me how wonderful and limitless the world of science is. Numerous discussions with my mentors through face-to-face talks, Skype talks, and email exchanges helped me countless times, and I definitely could not have finished this project without their support and guidance. When I faced a difficulty using DESMO-J, I sent a help email to a researcher from DESMO-J team, Department of Computer Science, University of Hamburg, and he aided me in my time of need by helping find the right API for the Poisson distribution. Without him, my project would have faced an unsolvable problem. I would also like to thank my parents for their endless encouragement and support towards my project. Whether it was driving eight hour to and from the university or listening to my outlandish ideas, my parents were essential to my success.

REFERENCES

[1] Bora Ristic, Kaveh Madani, and Zen Makuch, “The Water Footprint of Data Centers,” Sustainability 2015, 7, 11260-11284.

[2] Y. Sverdlik, “Google to build and lease data centers in big cloud expansion,” Data Center Knowledge, April 2016.

[3] Y. Sverdlik, “Here’s how much energy all US data centers consume” Data Center Knowledge, July 2016.

[4] Y. Sverdlik, “Here’s how much water all US data centers consume” Data Center Knowledge, July 2016.

[5] I. Goiri, W. Katsak, K. Le, T. D. Nguyen, and R. Bianchini, “Parasol and greenswitch: managing datacenters powered by renewable energy,” in ASPLOS, 2013.

[6] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, “Cutting the electric bill for internet-scale systems,” in SIGCOMM, 2009.

[7] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, “It’s not easy being green,” SIGCOMM Comput. Commun. Rev., 2012.

[8] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, and T. D. Nguyen, “Reducing electricity cost through virtual machine placement in high performance computing clouds,” in SuperComputing, 2011.

[9] R. Singh, D. Irwin, P. Shenoy, and K. K. Ramakrishnan, “Yank: Enabling green data centers to pull the plug,” in NSDI, 2013

[10] AT&T Water Management, http://about.att.com/content/csr/home/issue-brief-builder/environment/water-management.html

[11] Mentor, “Optimizing Water Efficiency in Distributed Data Centers,” International Conference on Cloud and Green Computing (CGC), 2013.

[12] Mentor, A. Gandhi, and Mentor, “Minimizing Electricity Cost for Geo-Distributed Interactive Services with Tail Latency Constraint,” International Green and Sustainable Computing Conference (IGSC), 2016.

[13] Mentor, Mentor, N. Pissinou, H. Mahmud, and A. V. Vasilakos, “Distributed Resource Management in Data Center with Temperature Constraint,” International Green Computing Conference (IGCC), 2013.

[14] Mentor, K. Ahmed, Mentor, and G. Quan, “Exploiting Temporal Diversity of Water Efficiency to Make Data Center Less ‘Thirsty’,” USENIX International Conference on Autonomic Computing (ICAC, special track on Self-Aware Cyber-Physical Systems), 2014.

[15] K. Le, R. Bianchini, T. D. Nguyen, O. Bilgir, and M. Martonosi, “Capping the brown energy consumption of internet services at low cost,” in IGCC, 2010.

[16] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, “Greening geographical load balancing,” in SIGMETRICS, 2011.

[17] L. Rao, X. Liu, L. Xie, and W. Liu, “Reducing electricity cost: Optimization of distributed Internet data centers in a multi-electricity-market environment,” in INFOCOM, 2010.

[18] Y. Zhang, Y. Wang, and X. Wang, “Electricity bill capping for cloud-scale data centers that impact the power markets,” in ICPP, 2012.

[19] Y. Zhang, Y. Wang, and X. Wang, “Greenware: Greening cloud-scale data centers to maximize the use of renewable energy,” in Middleware, 2011.

[20] D. S. Palasamudram, R. K. Sitaraman, B. Urgaonkar, and R. Urgaonkar, “Using batteries to reduce the power costs of internet-scale distributed networks,” in SoCC, 2012.

[21] 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 NSDI, 2015.

[22] C.-C. Hung, L. Golubchik, and M. Yu, “Scheduling jobs across geodistributed datacenters,” in SoCC, (New York, NY, USA), 2015.

[23] M. E. Haque, Y. h. Eom, Y. He, S. Elnikety, R. Bianchini, and K. S. McKinley, “Few-to-many: Incremental parallelism for reducing tail latency in interactive services,” in ASPLOS, 2015.

[24] “Google transparency report,” http://www.google.com/transparencyreport/traffic/explorer.

[25] DESMO-J Tutorial, http://desmoj.sourceforge.net/tutorial/overview/0.html, Department of Computer Science, University of Hamburg, Germany.

[26] Energy Calculator for Computers, https://www.eu-energystar.org/calculator.htm, European Union Energy Star.

[27] "Data Center Talk," http://www.datacentertalk.com/forum/showthread.php?t=31192.

[28] Zhou Feng, Tian Xin, Ma Guoyuan, "Investigation into the energy consumption of a data center with a thermosyphon heat exchanger," https://link.springer.com/content/pdf/10.1007/s11434-011-4500-5.pdf.

[29] Ambrose McNevin, "APAC data center survey reveals high PUE figures across the region," http://www.datacenterdynamics.com/news/apac-data-center-survey-reveals-high-pue-figures-across-the-region/75116.fullarticle.

[30] "Prineville, OR Data Center," https://www.facebook.com/PrinevilleDataCenter/app/399244020173259

[31] "Forest City, NC Data Center," https://www.facebook.com/ForestCityDataCenter/app/288655784601722

[32] Yunshui Chen, "AIRSYS A data centre cooling perspective from China," https://www.datacentreworld.com/__media/Future-D1-10-50-11-15-Yunshui-Chen-v2.pdf

[33] NTT Communications, "Hong Kong Financial Data Center," http://www.hk.ntt.com/content/dam/nttcom/hk/pdf/leaflet/Hong_Kong_Financial_Data_Center_Brochure_EN.pdf

[34] Ying Qin, Elizabeth Curmi, Grant M. Kopec, Julian M. Allwood, Keith S. Richards, "China's energy-water nexus – assessment of the energy sector's compliance with the "3 Red Lines" industrial water policy," http://www.sciencedirect.com/science/article/pii/S0301421515001196

[35] Vincent Tidwell, Barbie Moreland, "Mapping water consumption for energy production around the Pacific Rim," http://iopscience.iop.org/article/10.1088/1748-9326/11/9/094008/pdf

[36] Michael Patterson, "WP#35 - WATER USAGE EFFECTIVENESS (WUE): A GREEN GRID DATA CENTER SUSTAINABILITY METRIC," https://www.thegreengrid.org/en/resources/library-and-tools/238-Water-Usage-Effectiveness-%28WUE%29%3A-A-Green-Grid-Data-Center-Sustainability-Metric-