ROBUST GEOGRAPHICAL LOAD BALANCING FOR SUSTAINABLE DATA CENTERS

Tianyi Chen, Yu Zhang*, Xin Wang†, and Georgios B. Giannakis**

*Dept. of ECE and DTC, University of Minnesota, 55455, USA
†Dept. of CSE and Key Lab for Inf. Sci. of EMW (MoE), Fudan University, 200433, China

ABSTRACT

A systematic framework is put forth in this paper to integrate renewable energy sources (RES), distributed storage units, cooling facilities, as well as dynamic pricing into the workload and energy management tasks for a data center network. To cope with RES uncertainty, the resource allocation task is formulated as a robust optimization problem minimizing the worst-case net cost. The resulting problem is reformulated as a convex program, and then solved in a distributed fashion using the dual decomposition approach. Numerical tests demonstrate the performance gain of the proposed approach over the existing alternative.

Index Terms— Geo-distributed data centers, renewable energy, smart grid, cloud computing, robust optimization.

1. INTRODUCTION

In order to reduce the electricity cost of data centers (DC), considerable efforts from both industry and academia have been made over the last decade [1]. Carbon emission concerns along with the large energy consumed by DCs challenge their sustainability [2]. Exploiting renewable energy sources (RES) is clearly key to sustainable DC operation [3, 4]. Yet, high penetration of renewables unavoidably brings increased variability and uncertainty to the traditional power system. A major issue with renewable-integrated energy management is to account for its random and nondispatchable nature, which motivates the use of energy storage units [5]. Supply-side energy management with distributed storage units was considered for a homogeneous DC [6, 7], and geo-distributed DCs [8]. Taking advantage of RES, a two-time scale Lyapunov optimization technique was developed to control the energy supply in both ahead-of-time and real-time settings [5]. Note that most of prior works (e.g., [5, 7, 9]) assumed that RES generation is either independent and identically distributed (i.i.d.), or precisely known a-priori, which is not realistic in practice. Hence, how to properly deal with the RES uncertainty for DCs' daily operations is still an open problem.

In this paper, we consider the optimal workload and energy management for a cloud network consisting of multiple geo-distributed mapping nodes and DCs. Distinct from existing works, *distribution-free* uncertainty sets of the *unknown* renewable generation, as well as a two-way energy trading mechanism are introduced to account for the stochastic and nondispatchable nature of RES. Built on practical models, the resource allocation task is formulated as a robust optimization problem, which minimizes the system's worst-case net cost subject to DC operational constraints. Leveraging the problem structure, we show that it can be cast as a convex program. Capitalizing on the dual decomposition approach, an efficient distributed solver is developed. The proposed algorithm is guaranteed to yield the optimal strategy of robust workload and energy management, which also facilitates distributed implementations among the mapping nodes and DCs. Finally, extensive numerical tests with real data corroborate the merits of the proposed framework and approaches.

Notation. $\mathbb{R}$ for real numbers; $(\cdot)'$ stands for vector and matrix transposition; and $[a]^+ := \max\{a, 0\}$. Finally, the indicator function $\mathbb{1}_{(A)}$ takes value 1 when the event A happens, and 0 otherwise.

2. SYSTEM MODELS

Consider a network with geographically distributed mapping nodes $\mathcal{J} := \{1, 2, \ldots, J\}$, and DCs $\mathcal{I} := \{1, 2, \ldots, I\}$, over a discrete-time scheduling horizon $\mathcal{T} := \{1, \ldots, T\}$. Mapping nodes first collect data requests from nearby areas, and then distribute them to different DCs. Each DC has three subsystems: a cooling (heat dissipation) subsystem, an IT subsystem, and a power supply subsystem.

2.1. Network and workload models

In general, DC workloads are either delay-sensitive ('must-serve') or delay-tolerant [10]. For 'must-serve' workloads, let A_j^t and a_{ji}^t denote the rate of service requests arriving at mapping node j, and the one from node j to DC i in slot t, respectively. For delay-tolerant workloads, let $\mathcal{Q}_j$ denote the jobs collected by node j, and $\mathcal{Q} := \bigcup_{j=1}^J \mathcal{Q}_j$ with $\mathcal{Q}_i \bigcap \mathcal{Q}_j = \emptyset, \forall i \neq j$, representing the set of all delay-tolerant jobs. The qth delay-tolerant job can be specified by its total demand W_q and active interval $\mathcal{T}_q := \{S_q, \ldots, E_q\}$. Let $\tilde{w}_{i,q}^t$ and $w_{i,q}^t$ denote the amount of qth delay-tolerant job routed from its original mapping node to DC i, and the amount being processed by DC i in slot t, respectively; and L_{ji}^t denote the link bandwidth from node j to DC i at time t. As shown in Fig. 1, these quantities must satisfy the following constraints:

$$\sum\nolimits_{i=1}^{I} a_{ji}^t = A_j^t,\ \forall j \in \mathcal{J},\ t \in \mathcal{T} \tag{1}$$

$$\sum\nolimits_{t=S_q}^{E_q} \sum\nolimits_{i=1}^{I} \tilde{w}_{i,q}^t = W_q,\ \ \forall q \in \mathcal{Q} \tag{2}$$

$$a_{ji}^t + \sum\nolimits_{q \in \mathcal{Q}_j} \tilde{w}_{i,q}^t \leq L_{ji}^t,\ \forall i \in \mathcal{I},\ j \in \mathcal{J},\ t \in \mathcal{T} \tag{3}$$

where (1) ensures that 'must-serve' workloads are dispatched once arrived; (2) requires routing each delay-tolerant job before its deadline; and (3) captures the bandwidth limitation of data transfer.

Per DC, 'must-serve' workloads are processed immediately, while delay-tolerant workloads are deferrable. DC operations of

Work in this paper was supported by NSF 1509040, 1508993, 1509005, 1423316, 1442686 and 1202135; the China Recruitment Program of Global Young Experts, the Program for New Century Excellent Talents in University, the Innovation Program of Shanghai Municipal Education Commission.

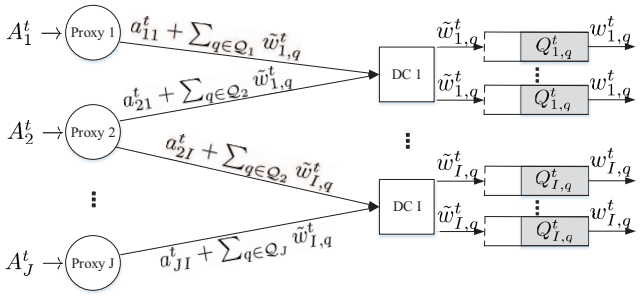

Fig. 1. A workload distribution diagram.

delay-tolerant workloads can be described as

$$\sum_{\tau=S_q}^{E_q} \tilde{w}_{i,q}^{\tau} = \sum_{\tau=S_q}^{E_q} w_{i,q}^{\tau},\ \forall i \in \mathcal{I}, q \in \mathcal{Q} \tag{4a}$$

$$\sum_{\tau=S_q}^{t} \tilde{w}_{i,q}^{\tau} \geq \sum_{\tau=S_q}^{t} w_{i,q}^{\tau},\ \forall i \in \mathcal{I}, q \in \mathcal{Q}, t \in [S_q, E_q - 1] \tag{4b}$$

where (4a) adheres to the deadline completion requirements, while (4b) entails the causality of delay-tolerant workloads. The total IT demand of DC i in slot t, is thus given by

$$d_i^t = \sum\nolimits_{j \in \mathcal{J}} a_{ji}^t + \sum\nolimits_{q \in \mathcal{Q}} w_{i,q}^t,\ \forall t \in \mathcal{T}. \tag{5}$$

2.2. Power demand and supply models

Let m_i^t denote the number of active servers in DC i at time t that satisfies

$$\underline{M}_i \leq m_i^t \leq \overline{M}_i, \quad 0 \leq d_i^t \leq m_i^t D_i \tag{6}$$

where $\underline{M}_i, \overline{M}_i$ stand for the minimum and maximum number of homogeneous servers, and D_i is the server capacity[1]. With each server running at a speed of $d_i^t/(m_i^t D_i)$, the total power consumption is

$$P_{i,IT}(d_i^t, m_i^t) = \frac{\varrho {d_i^t}^2}{m_i^t D_i^2} + (1 - \varrho) m_i^t$$

where $1 - \varrho$ denotes the power consumed in the idle state [12].

Along with the increasing density of IT equipment in DCs, a considerable amount of electricity is consumed by the cooling system that generally operates in two modes [7, 13]: *outside-air* and *chilled-water* cooling. Due to different efficiencies and capacities of the two cooling approaches, for a given $P_{i,IT}$, there is an optimal allocation between outside-air cooling and chiller cooling. It turns out that the cooling consumption minimization admits a closed-form solution [7]

$$F_i^t(P_{i,IT}) = \begin{cases} \kappa_i^t (P_{i,IT})^3, & P_{i,IT} \leq P_{i,s}^t \\ \kappa_i^t (P_{i,s}^t)^3 + \gamma (P_{i,IT} - P_{i,s}^t), & P_{i,IT} > P_{i,s}^t \end{cases}$$

where κ_i^t and $P_{i,s}^t$ are temperature-dependent parameters, and constant γ is the cooling coefficient of chilled-water cooling. For notational convenience, let $P_i^t(d_i^t, m_i^t) := F_i^t(d_i^t, m_i^t) + P_{i,IT}(d_i^t, m_i^t)$, which is jointly convex in $\{d_i^t, m_i^t\}$.

We consider each DC to be supplied by a RES-integrated microgrid consisting of a conventional generator (CG) (e.g., fuel generator), an on-site renewable generator (RG) (e.g., wind or solar), and an energy storage unit (e.g., battery).

[1]Since the number of servers is very large, m_i^t can be relaxed to be a positive real number for simplicity [11].

Let $P_{i,g}^t$ denote the energy output of the CG in DC i per slot t, which is upper bounded by $\overline{P}_{i,g}$; that is,

$$0 \leq P_{i,g}^t \leq \overline{P}_{i,g},\ \forall i \in \mathcal{I}, t \in \mathcal{T}. \tag{7}$$

The change of CG energy output in two consecutive slots is bounded by the following so-termed ramping constraints:

$$P_{i,g}^t - P_{i,g}^{t-1} \leq R_i^{\rm up},\ \ P_{i,g}^{t-1} - P_{i,g}^t \leq R_i^{\rm dw},\ \forall i \in \mathcal{I}, t \in \mathcal{T} \tag{8}$$

where $R_i^{\rm up}$ and $R_i^{\rm dw}$ are the ramping-up and -down rates of CG.

Let $P_{i,b}^t$ denote the power delivered to (or drawn from) the storage unit in DC i at slot t, which amounts to either charging ($P_{i,b}^t > 0$) or discharging ($P_{i,b}^t < 0$). Let C_i^0 and C_i^t denote the initial amount of stored energy and the state of charge (SoC) of the storage unit in DC i at the beginning of time slot t. Each unit has a finite capacity $\overline{C}_i$ as well as a minimum level $\underline{C}_i$. In short, the energy storage unit can be compactly described as

$$\underline{C}_i \leq C_i^t \leq \overline{C}_i,\ \ \forall i \in \mathcal{I}, t \in \mathcal{T} \tag{9}$$

$$C_i^{t+1} = C_i^t + P_{i,b}^t,\ \ \forall i \in \mathcal{I}, t \in \mathcal{T} \tag{10}$$

$$\underline{P}_{i,b} \leq P_{i,b}^t \leq \overline{P}_{i,b},\ \ \forall i \in \mathcal{I}, t \in \mathcal{T} \tag{11}$$

where the bounds of (dis-)charging amount $\underline{P}_{i,b} < 0$ and $\overline{P}_{i,b} > 0$ are dictated by physical limits.

Consider now the RES vector $\mathbf{e}_i := [E_i^1, \ldots, E_i^T]'$ generated at DC i across all slots. Due to the unpredictable and intermittent nature of RES, $\mathbf{e}_i$ is unknown a priori. The proposed method of postulating an uncertainty region provides the decision maker with ranges instead of point forecasts, which is essentially distribution-free and *robust* to prediction errors. The actual RES generation $\mathbf{e}_i$ is assumed to lie in a polyhedral uncertainty set $\mathcal{E}_i$ (see also [14]):

$$\mathcal{E}_i := \left\{ \mathbf{e}_i \mid \underline{E}_i^t \leq E_i^t \leq \overline{E}_i^t,\ E_i^{\min} \leq \sum\nolimits_{t \in \mathcal{T}} E_i^t \leq E_i^{\max} \right\} \tag{12}$$

where $\underline{E}_i^t$ ($\overline{E}_i^t$) denotes a lower (upper) bound on E_i^t, and the total renewables at DC i over the scheduling horizon is bounded by $E_i^{\min}$ and $E_i^{\max}$.

2.3. Cost-revenue model

In addition to the internal energy resources (namely, CG, RG, and storage unit), DCs can resort to the main grid market in an on-demand manner. Suppose that the energy can be purchased from the wholesale electricity market around DC i in period t at price α_i^t, while the energy is sold at price β_i^t. Notwithstanding, we shall always set $\alpha_i^t \geq \beta_i^t$ to avoid less relevant buy-and-sell activities of the DC for profit. Let $\tilde{P}_i^t := P_i^t + P_{i,b}^t \mathbb{1}_{(P_{i,b}^t > 0)}$ denote the total energy consumption of DC i per slot t, and $S_i^t := P_{i,g}^t + E_i^t - P_{i,b}^t \mathbb{1}_{(P_{i,b}^t < 0)}$ the total energy supply in DC i per slot t. For DC i, the *worst-case transaction cost* for the whole scheduling horizon is defined as $G_i(\{\tilde{P}_i^t\}, \{S_i^t\}) := \max_{\mathbf{e}_i \in \mathcal{E}_i} \sum_{t=1}^T \alpha_i^t [\tilde{P}_i^t - S_i^t]^+ - \beta_i^t [S_i^t - \tilde{P}_i^t]^+$. With $\psi_i^t := (\alpha_i^t - \beta_i^t)/2$, $\phi_i^t := (\alpha_i^t + \beta_i^t)/2$, and $R_i^t = P_i^t + P_{i,b}^t - P_{i,g}^t$, we rewrite $G_i(\{\tilde{P}_i^t\}, \{S_i^t\})$ as

$$G_i(\{R_i^t\}) = \max_{\mathbf{e}_i \in \mathcal{E}_i} \sum_{t=1}^T \left(\psi_i^t |R_i^t - E_i^t| + \phi_i^t (R_i^t - E_i^t) \right). \tag{13}$$

In addition, let $G_{C_i}(P_{i,g}^t)$ denote the cost of CG at DC i in slot t, which is convex piecewise linear or smooth quadratic. The revenue earned per slot t can be modeled as a concave function, e.g., $U_q^t(w_{i,q}^t) = u_q^t w_{i,q}^t$, where u_q^t is the revenue per unit of qth workloads at time t.

3. ROBUST GEOGRAPHICAL LOAD BALANCING

Based on the practical models in Section II, we pursue in this section a robust workload and energy management approach for the considered DC network. Over the scheduling horizon $\mathcal{T}$, the system operator per mapping node performs an (e.g. hour-) ahead-of-time schedule to optimize workload routing $\{a_{ji}^t, \tilde{w}_{i,q}^t\}$, while the system operator in each DC optimizes scheduling of servers and workloads $\{m_i^t, w_{i,q}^t\}$, CG generation $\{P_{i,g}^t\}$, and battery (dis-)charging energy $\{P_{i,b}^t\}$. With $\mathbf{x}$ collecting optimization variables $\{a_{ji}^t, w_{i,q}^t, \tilde{w}_{i,q}^t, d_i^t, m_i^t, P_{i,g}^t, P_{i,b}^t, R_i^t, C_i^t\}$, the system operator wants to solve the following problem:

$$\min_{\mathbf{x}} \sum_{i=1}^{I} G_i(\{R_i^t\}) + \sum_{t=1}^{T}\sum_{i=1}^{I}\left(G_{C_i}(P_{i,g}^t) - \sum_{q\in\mathcal{Q}} U_q^t(w_{i,q}^t)\right) \tag{14a}$$

$$\text{s.t.} \quad R_i^t \geq P_i^t + P_{i,b}^t - P_{i,g}^t,\ \forall i \in \mathcal{I}, t \in \mathcal{T} \tag{14b}$$

$$(1) - (11). \tag{14c}$$

It is worth mentioning that thanks to the worst-case transaction cost $G_i(\{R_i^t\})$, the RES induced randomness has been eliminated; thus, (14) contains only deterministic variables. Since the objective (14a) is monotonically increasing with R_i^t, it is easy to see that (14b) is always binding at the optimal solution $\mathbf{x}^*$, which entails the convexity of (14). However, the objective (14a) is to minimize a pointwise maximum function, which is generally non-differentiable.

3.1. Lagrange dual approach

Notice that constraints (1)–(3) and (14b) couple variables across mapping nodes, DCs, workloads, and the RES, so a system operator over the entire network is essential to collect all the information and solve the problem in a centralized fashion, which may not be feasible in an Internet-scale network [15]. However, since (14) is a convex problem, a Lagrange dual approach can be developed to efficiently find its optimal dual solution with zero duality gap in a decentralized manner [16]. Letting $\boldsymbol{\varpi} := \{\lambda_{i,q}^t, \nu_i^t, \pi_i^t\}$ collect all the Lagrange multipliers associated with the constraints (4)–(5) and (14b)[2], the partial Lagrangian function of (14) is

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \boldsymbol{\varpi}) :=& \sum_{i=1}^{I}\left[G_i(\{R_i^t\}) + \sum_{t=1}^{T}\left(G_{C_i}(P_{i,g}^t) - \sum_{q\in\mathcal{Q}} U_q^t(w_{i,q}^t)\right)\right] \\ &+ \sum_{i=1}^{I}\sum_{t=1}^{T}\nu_i^t\left(d_i^t - \sum\nolimits_{j=1}^{J} a_{ji}^t - \sum\nolimits_{q\in\mathcal{Q}} w_{i,q}^t\right) \\ &+ \sum_{i=1}^{I}\sum_{q\in\mathcal{Q}}\sum_{t=1}^{T}\lambda_{i,q}^t\left(\sum\nolimits_{\tau=S_q}^{t} w_{i,q}^\tau - \sum\nolimits_{\tau=S_q}^{t}\tilde{w}_{i,q}^\tau\right) \\ &+ \sum_{i=1}^{I}\sum_{t=1}^{T}\pi_i^t\left(P_i^t + P_{i,b}^t - P_{i,g}^t - R_i^t\right).\end{aligned}$$

With $\mathcal{X}$ denoting the set given by constraints (1)–(3), and (6)–(11), the dual function is thus $\mathcal{D}(\boldsymbol{\varpi}) := \min_{\mathbf{x}\in\mathcal{X}}\ \mathcal{L}(\mathbf{x}, \boldsymbol{\varpi})$, and the dual problem of (14) is

$$\begin{aligned}&\max_{\boldsymbol{\varpi}}\ \mathcal{D}(\{\pi_i^t\}, \{\lambda_{i,q}^t\}, \{\nu_i^t\}) \\ &\text{s.t.}\ \ \pi_i^t \geq 0,\ \nu_i^t \in \mathbb{R},\ \forall i, t \\ &\qquad \lambda_{i,q}^{E_q} \in \mathbb{R},\ \lambda_{i,q}^t \geq 0,\ \forall i, q, t \in [S_q, E_q - 1].\end{aligned} \tag{15}$$

[2]For notational convenience, let $\lambda_{i,q}^t = 0, \forall i, q \in \mathcal{Q}, t \notin \mathcal{T}_q$.

For the dual problem (15), the projected subgradient method can be employed to obtain the optimal $\boldsymbol{\varpi}^*$, namely

$$\boldsymbol{\varpi}(k+1) = \text{proj}\left(\boldsymbol{\varpi}(k) + \mu g_{\boldsymbol{\varpi}}(k)\right) \tag{16}$$

where proj$(\cdot)$ is the projection operator to the feasible set of $\boldsymbol{\varpi}$; k is the iteration index; $\mu > 0$ is a constant stepsize; and $g_{\boldsymbol{\varpi}}(k) := \{g_{\pi_i^t}(k), g_{\lambda_{i,q}^t}(k), g_{\nu_i^t}(k)\}$ are the subgradients of $\mathcal{D}(\boldsymbol{\varpi})$ with respect to the Lagrange multipliers. Specifically, we have

$$g_{\pi_i^t}(k) = P_i^t(k) + P_{i,b}^t(k) - P_{i,g}^t(k) - R_i^t(k) \tag{17a}$$

$$g_{\lambda_{i,q}^t}(k) = \sum\nolimits_{\tau=S_q}^{t} w_{i,q}^\tau(k) - \sum\nolimits_{\tau=S_q}^{t}\tilde{w}_{i,q}^\tau(k) \tag{17b}$$

$$g_{\nu_i^t}(k) = d_i^t(k) - \sum\nolimits_{j=1}^{J} a_{ji}^t(k) - \sum\nolimits_{q\in\mathcal{Q}} w_{i,q}^t(k) \tag{17c}$$

where primal variables $\mathbf{x}(k)$ can be obtained by

$$\min_{\{a_{ji}^t, \tilde{w}_{i,q}^t\}} \sum_{t=1}^{T}\sum_{i=1}^{I}\left[-a_{ji}^t\nu_i^t(k) - \sum\nolimits_{q\in\mathcal{Q}_j}\tilde{w}_{i,q}^t\sum\nolimits_{\tau=t}^{T}\lambda_{i,q}^\tau(k)\right] \quad \text{s.t.}\ (1)-(3) \tag{18}$$

$$\min_{\{0\leq w_{i,q}^t \leq W_q\}} \sum_{t=1}^{T}\left[w_{i,q}^t\left(\sum\nolimits_{\tau=t}^{T}\lambda_{i,q}^\tau(k) - \nu_i^t(k)\right) - U_q^t(w_{i,q}^t)\right] \tag{19}$$

and

$$\begin{aligned}\min_{\substack{\{R_i^t, m_i^t,\\ P_{i,b}^t, P_{i,g}^t, d_i^t\}}}\ & G_i(\{R_i^t\}) + \sum_{t=1}^{T}\left[\nu_i^t(k)d_i^t + G_{C_i}(P_{i,g}^t)\right. \\ &\left. + \pi_i^t(k)(P_i^t + P_{i,b}^t - P_{i,g}^t - R_i^t)\right] \quad \text{s.t.}\ (6)-(11).\end{aligned} \tag{20}$$

The subproblems (18) and (19) are linear programs (LPs), which can be optimally solved using off-the-shelf algorithms. However, since $G_i(\{R_i^t\})$ is non-differentiable due to the absolute value operator and the maximization over $\mathbf{e}_i \in \mathcal{E}_i$, (20) still challenges existing solvers. To address this, consider splitting (20) into two subproblems, namely

$$\min_{\substack{\{m_i^t, P_{i,b}^t,\\ P_{i,g}^t, d_i^t\}}} \sum_{t=1}^{T}[\pi_i^t(k)(P_i^t + P_{i,b}^t - P_{i,g}^t) + \nu_i^t(k)d_i^t + G_{C_i}(P_{i,g}^t)] \quad \text{s.t.}\ (6)-(11) \tag{21}$$

and

$$\min_{\{\underline{R}_i \leq R_i^t \leq \overline{R}_i\}}\ G_i(\{R_i^t\}) - \sum_{t=1}^{T}\pi_i^t(k)R_i^t. \tag{22}$$

where $\underline{R}_i$ and $\overline{R}_i$ are lower and upper bounds of the right hand side of (14b). Depending on the function $G_{C_i}(P_{i,g}^t)$, subproblem (21) is either an LP or a quadratic program, which is efficiently solvable. And for nonsmooth subproblems (22), a standard subgradient iteration can be employed to obtain the optimal solution as

$$R_i^t(\ell+1) = R_i^t(\ell) - \eta_\ell g_{R_i^t}(\ell),\ \forall t \in \mathcal{T} \tag{23}$$

where ℓ denotes iteration index, and $\{\eta_\ell\}$ is an appropriate stepsize sequence. The partial subgradient of $G_i(\{R_i^t\})$ with respect to R_i^t can be obtained as

$$g_{R_i^t}(\ell) := \frac{\partial G_i(\{R_i^t\})}{\partial R_i^t} = \begin{cases}\alpha_i^t - \pi_i^t(k), & \text{if } R_i^t(\ell) \geq E_i^{t*}(\ell) \\ \beta_i^t - \pi_i^t(k), & \text{if } R_i^t(\ell) < E_i^{t*}(\ell)\end{cases}$$

Algorithm 1 Distributed workload and energy management

1: **Initialize:** Choose a proper $\boldsymbol{\varpi}(0)$ and stepsize μ
2: **repeat** $k = 0, 1, 2 \ldots$
3: Each DC solves (19) and (21)–(22) separately to obtain $\{w_{i,q}^t(k), R_i^t(k), m_i^t(k), P_{i,b}^t(k), P_{i,g}^t(k), d_i^t(k)\}$
4: Each mapping node solves (18) and sends $\{a_{ji}^t(k), \tilde{w}_{i,q}^t(k)\}$ to each DC
5: DCs update $\boldsymbol{\varpi}(k)$ via (16) and send them to mapping nodes
6: Run averages to approximate primal variables via

$$\overline{\mathbf{x}}(k) = \frac{1}{k}\mathbf{x}(k-1) + \frac{k-1}{k}\overline{\mathbf{x}}(k-1)$$

7: **until** Convergence

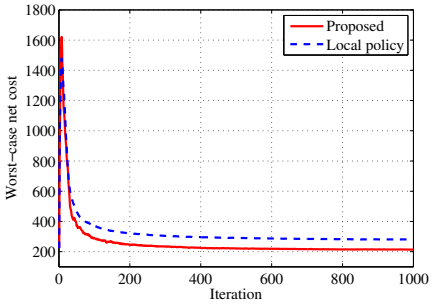

Fig. 2. Comparison of worst-case net costs.

where $\mathbf{e}_i^*(\ell) := [E_i^{1*}(\ell), \ldots, E_i^{T*}(\ell)]'$ for the given $\{R_i^t(\ell)\}$ is found by solving [cf. (12) and (13)]

$$\max_{\mathbf{e}_i \in \mathcal{E}_i} \sum_{t=1}^{T} \left(\psi_i^t |R_i^t(\ell) - E_i^t| + \phi_i^t (R_i^t(\ell) - E_i^t)\right). \qquad (24)$$

Under the condition $\alpha_i^t \geq \beta_i^t, \forall t \in \mathcal{T}$, problem (24) is essentially convex maximization over a polyhedron, which is generally NP-hard. Fortunately, the globally optimal solution is attainable at the extreme points of $\mathcal{E}_i$ [17, Sec. 2.4]. Leveraging the polyhedral structure of $\mathcal{E}_i$, we adopt an efficient vertex enumerating algorithm to obtain $\mathbf{e}_i^*$ efficiently.

3.2. Optimality and distributed implementation

For the subgradient iterations (23), if a diminishing stepsize satisfying (i) $\sum_{\ell=0}^{\infty} \eta_\ell = \infty$, and (ii) $\sum_{\ell=0}^{\infty} \eta_\ell^2 < \infty$ is adopted, the sequence (16) converges as $\ell \to \infty$ to the optimal $\{R_i^t(k)^*\}$. Regarding the constant stepsize μ in (16), the subgradient iterations will converge to a neighborhood of the optimal solution $\boldsymbol{\varpi}^*$. The size of the neighborhood is proportional to the stepsize μ [17]. Since the objective of (14) is not strictly convex, running averages of the primal sequence $\{\mathbf{x}(k)\}$ can be used to recover the optimal primal solutions. It is also worth noting that the considered robust geographical load balancing facilitates a distributed implementation, where optimization tasks are distributed among mapping nodes and individual DCs; see Algorithm 1.

4. NUMERICAL EVALUATION

The DC network includes six DCs and six mapping nodes uniformly located in the eastern, central, and western US. The time horizon spans $T = 12$ hours, corresponding to the interval 1PM–12AM in Eastern Time Zone. Notice that we use the Eastern Time Zone for time-keeping, so the peaks of workload demands, RES and prices

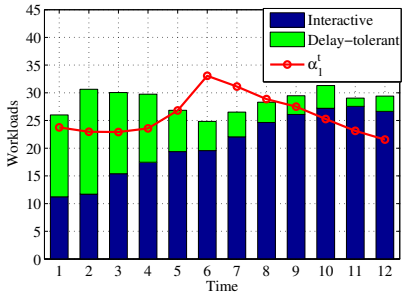

Fig. 3. Optimal workloads schedule of DC 1.

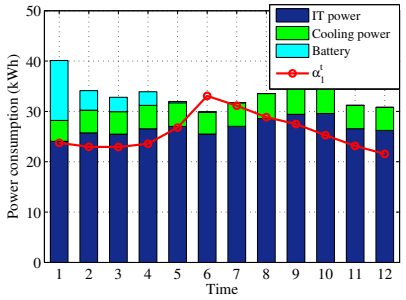

Fig. 4. Optimal power consumption schedule of DC 1.

are different in the three areas. Finally, a robust *local policy* is introduced as a benchmark, where workloads received by each mapping node are distributed only to its nearest DC. The specific parameter configurations are omitted due to limited space.

In Fig. 2, the proposed algorithm is compared with the local policy in terms of their worst-case net costs. Within 500 iterations, the proposed algorithm converges to a worst-case net cost 25% lower than that of the local policy. This is because the proposed algorithm takes both spatial and temporal variations into account. For instance, mapping nodes can intelligently route workloads to a remote DC where the system demand is lower, RES availability is higher, or, the local energy price is more affordable.

Fig. 3 depicts the optimal workloads schedule of DC 1. One observation is that real-time IT demand closely reflects the instantaneous energy purchase price α_1^t. Specifically, the proposed method tends to schedule more workloads when purchase price α_1^t is low (1PM-5PM), and vice versa. Moreover, flexible delay-tolerant workloads are more likely to be processed when the 'must-serve' demand is low, or, when the purchase price is low. This corroborates the merit of our proposed algorithm in "smoothing" the IT demand curve.

The optimal power consumption schedules of DC 1 is depicted in Fig. 4, where less power is consumed when α_1^t is higher (6PM). Using combined cooling sources, the average cooling coefficient of the proposed algorithm is around 0.17, which is more efficient than the simple chilled-water cooling with a constant coefficient $\gamma = 0.2$. With the goal of mitigating the high variability of RES, batteries are encouraged to charge when the worst-case renewable generations are high and the energy prices are low (1PM-4PM).

5. CONCLUSIONS

Robust ahead-of-time workload and energy management for green DCs was considered in this paper. Taking into account the spatio-temporal variation of workloads, renewables and electricity market prices, a resource allocation problem was formulated to minimize the system net cost including the network operational cost and the worst-case energy transaction cost. Relying on the strong duality of the convex reformulation, a Lagrange dual based distributed solver was developed to yield the optimal solution.

6. REFERENCES

[1] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for Internet-scale systems," in *Proc. of ACM SIGCOMM*, Barcelona, Spain, Aug. 2009, vol. 39, pp. 123–134.

[2] P. X. Gao, A. R Curtis, B. Wong, and S. Keshav, "It's not easy being green," in *Proc. of ACM SIGCOMM*, Helsinki, Finland, Aug. 2012, vol. 42, pp. 211–222.

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

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

[5] W. Deng, F. Liu, H. Jin, C. Wu, and X. Liu, "Multigreen: Cost-minimizing multi-source data center power supply with online control," in *Proc. of ACM Intl. Conf. on Future Energy systems*, Berkeley, CA, May 2013, pp. 149–160.

[6] R. Urgaonkar, B. Urgaonkar, M. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *Proc. of ACM SIGMETRICS*, San Jose, CA, Jun. 2011, pp. 221–232.

[7] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *Proc. of ACM SIGMETRICS*, London, UK, Jun. 2012, vol. 40, pp. 175–186.

[8] Y. Guo and Y. Fang, "Electricity cost saving strategy in data centers by using energy storage," *IEEE Trans. Parallel and Distrib. Syst.*, vol. 24, no. 6, pp. 1149–1160, Jun. 2013.

[9] 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 and Distrib. Syst.*, vol. 25, no. 8, pp. 2030–2042, Aug. 2014.

[10] 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. of INFOCOM*, Orlando, FL, Mar. 2012, pp. 1431–1439.

[11] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Trans. Networking*, vol. 21, no. 5, pp. 1378–1391, Oct. 2013.

[12] D. Xu and X. Liu, "Geographic through filling for Internet datacenters," in *Proc. of INFOCOM*, Orlando, FL, Mar. 2012, pp. 2881–2885.

[13] Active Power, "Data center thermal runaway. a review of cooling challenges in high density mission critical environments," *White Paper*, 2007.

[14] Y. Zhang, N. Gatsis, and G. B. Giannakis, "Robust energy management for microgrids with high-penetration renewables," *IEEE Trans. Sustain. Energy*, vol. 4, no. 4, pp. 944–953, Oct. 2013.

[15] P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford, "Donar: Decentralized server selection for cloud services," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 231–242, Aug. 2011.

[16] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[17] D. P. Bertsekas, *Convex Optimization Theory*, Athena Scientific, Belmont, MA, 2009.