Toward Low-Cost Workload Distribution for Integrated Green Data Centers

---

Citation:
A. Kiani and N. Ansari, "Towards low-cost workload distribution for integrated green data centers," IEEE Communications Letters, vol. 19, no. 1, pp. 26–29, 2015.

URL:
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6953198

Towards Low-Cost Workload Distribution for Integrated Green Data Centers

Abbas Kiani *Student Member, IEEE*, Nirwan Ansari *Fellow, IEEE*

***Abstract*—This paper aims at maximizing the utilization of green energy and cutting the cost of electricity associated in provisioning computing services across a group of data centers. To this end, we propose the notion of green workload and green service rate, versus brown workload and brown service rate, respectively, to facilitate the separation of green energy utilization maximization and brown energy cost minimization problems. Accordingly, a workload distribution algorithm is designed such that the cost of electricity is reduced as compared to the existing workload distribution schemes.**

***Keywords*—Data centers, Cost of Electricity, Green energy**

I. Introduction

THE exponentially growing demand for online services that run on hundreds of thousands of servers spread across large data centers has significantly craved electric power usage [1]. Meeting such a demand in an environmentally friendly manner calls for innovations across different disciplines. Some efforts have been made to design energy efficient data centers in the past few years. New power management techniques have been developed to reduce not only carbon footprints but also the cost of electricity associated with data centers. Recently, renewable energy resources such as solar panels and wind turbines have been integrated into data centers, thereby promoting sustainability and green energy [2], [3]. However, there has been few research conducted considering a group of data centers [2], [4], [5]. Rao *et al.* [4] considered the electricity markets of Internet Data Centers (IDCs) spread across geographical diversity and studied the problem of minimizing the cost of electricity. They modeled and solved the minimization problem as the mixed-integer programming constrained with guaranteed quality of service. Moreover, Li *et al.* [5] proposed another solution based on mixed integer programming. The proposed model not only minimizes the electricity cost of geographically dispersed IDCs but also optimizes server on/off scheduling. While the main idea in [4], [5] is to cut the electricity costs of IDCs, these studies do not consider the potential advantages of integration of renewable resources. A workload distribution strategy is investigated in [2] by taking into account of the availability of renewable generators at different data centers. While the renewable energy and brown energy incur different costs and different environmental impacts, none of the existing schemes proposed for a group of data centers taps on the potential merits of the separation of green energy utilization maximization and brown energy cost minimization problems via decomposition of the workload served by green and brown energy, respectively. Moreover, the proposed optimization-based workload distribution in [2] is based under the assumption that local renewable generation is always less than the local power consumption. However, it may happen that the available renewable energy at a data center is adequate or even more than the power consumption. Thus, we tackle this shortcoming by separating the green energy utilization maximization and brown energy cost minimization problems. We propose a framework to provision a Green and ecOnomic wOrkload Distribution (GOOD) algorithm by integrating a group of data centers dispersed at different locations. In our algorithm, each data center utilizes the green energy as much as possible, and purchases brown energy only when the green energy generation is not adequate to serve all incoming workload. Figure 1 depicts the system model with the consideration of green and economical factors. We will address the following:

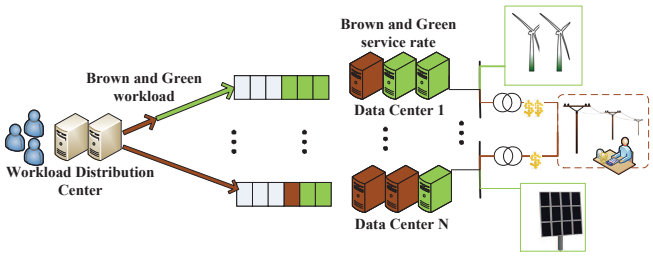

Fig. 1. System Model.

- **Workload Distribution Center**: As shown in Figure 1, one or a group of servers can serve as the workload distribution center [2]. The distribution center facilitates workload flexibility at the demand side. This center receives requests from all users and manages the distribution of the incoming workload to the geographically dispersed data centers based on the availability of green energy and the price of electricity. The center monitors the waiting requests at each data center, i.e., data centers queue lengths. Moreover, it predicts solar and wind energy generation at different data centers at different times.
- **Green versus Brown**: We propose the notion of green workload and green service rate versus brown workload and brown service rate, respectively. This concept facilitates the separation of green energy utilization maximization and brown energy cost minimization problems. In fact, the notion of green versus brown is employed

Authors are with the Advanced Networking Lab., Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, 07102 USA. (e-mail:ak628@njit.edu and Nirwan.Ansari@njit.edu)

to design and develop a green and low-cost workload distribution strategy.

II. System Model and Problem Formulation

Figure 1 shows the proposed system model, in which we consider a group of N data centers dispersed at different regions. To benefit from the energy efficiency and sustainability advantages of greening, each data center is integrated with a green power source such as wind turbine or solar panel. We assume that time is divided into several time slots at equal lengths. At each time slot t, a total number of $L(t)$ service requests, coming from all users, are received by the workload distribution center.

In our formulation, the total power consumption at each data center takes into account of the Base Load and Proportional Load [3],

$$Total\ Power\ Consumtion\ at\ data\ center\ i = m_i[P_{idle} + (E_{usage} - 1)P_{peak}] + m_i[(P_{peak} - P_{idle})U_i], \quad (1)$$

where the base load, $m_i[P_{idle} + (E_{usage} - 1)P_{peak}]$, indicates the power consumption even when all of the turned on servers are idle. The proportional load, $m_i[(P_{peak} - P_{idle})U_i]$, is the extra power consumption which is proportional to the CPU utilization of the servers, U_i, and accordingly to the workload. It is worth mentioning that both base and proportional loads are computed based on the number of switched on servers, m_i, idle power, P_{idle}, and average peak power of a single server, P_{peak}. Moreover, due to different energy efficiencies at different data centers, our definition of the total power consumption incorporates the Power Usage Effectiveness (PUE) ratio, E_{usage}, thereby amalgamating the power consumption at facility for cooling, lighting, etc [6].

It is assumed the data centers are offering a single class of Internet service. However, the problem formulation can be extended to multiple classes of service. Each incoming service request has to be processed within a deadline determined by the Service Level Agreement (SLA). Note that the service time of each service request depends on the queue length at its arrival. In other words, as the queue length increases, the queueing delay increases, and hence it takes more time to serve the request. Therefore, in this paper, we enforce the QoS by imposing an upper bound on each queue length.

A. Green energy problem formulation

Referring to the definition of power consumption in (1), the service rate at each data center, $\mu_i = m_i k$, can be expressed as a function of power. Note that k is the total number of requests that one server can handle per second. We define the *green service rate* as the achievable service rate at each data center powered by the available renewable energy. The green service rate at time slot t can be computed as $\mu_{g_i}(t) = m_{g_i}(t)k$, where $m_{g_i}(t)$ is the number of *green servers*, those servers that can be switched on and run at full utilization by green energy. The number of green servers is limited by the availability of green energy as well as the maximum number of servers at each location, i.e., $m_{g_i}(t) = \min(\lfloor \frac{W_i(t)}{P_{peak}E_{usage}} \rfloor, M_i)$, where $W_i(t)$ and M_i are the available green energy at the time slot and the maximum number of servers at data center i, respectively. Therefore, in order to consider the worst case scenario in our optimization framework, we define the green service rate as, $\mu_{g_i}(t) = \min((\frac{W_i(t)}{P_{peak}E_{usage}} - 1)k, M_i k)$. It is worth mentioning that the generated green energy can be predicted by the distribution center for different data centers at different time slots, i.e., by taking into account of weather dependency of green energy. Specifically, when the renewable source is wind turbine, the prediction can rely on the foremost forecasting techniques which are based on numeric weather prediction (NWP) of wind speed and power [7]. The prediction may include Very-Short Term Forecasting, Short Term Forecasting, Medium Term Forecasting and Long Term Forecasting techniques. Furthermore, if the case is solar generation, machine learning based prediction techniques [8] can be employed.

In addition, we define *green workload*, $\lambda_{g_i}(t)$, as the amount of workload forwarded to each data center based on the availability of renewable energy at that data center. Denote $\delta_i(t)$ as the queue length of the data center i at time t. We define $\delta_{g_i}(t)$ as the amount of queue length that can be served by green service rate, that is, $\delta_{g_i}(t) = \min(\delta_i(t), \mu_{g_i}(t))$.

B. Brown energy problem formulation

If green energy generation is lower than the required energy to serve all incoming workload, brown energy is used. Brown energy is considered as an additional resource to power on additional servers referred as the *brown servers*. Similar to green energy problem formulation, $\lfloor \frac{P_i(t)}{P_{peak}E_{usage}} \rfloor$ is the number of brown servers, $m_{b_i}(t)$, that are switched on and run at full utilization by brown energy and is upper bounded by $M_i - m_{g_i}(t)$. We define the *brown service rate* as the secured service rate powered by brown energy to ensure QoS requirements. Therefore, the brown service rate at a data center is established as, $\mu_{b_i}(t) = \min((\frac{P_i(t)}{P_{peak}E_{usage}} - 1)k, (M_i - m_{g_i}(t))k)$, where $P_i(t)$ indicates the total brown energy consumption at data center i at time t. Moreover, we define the *brown workload*, $\lambda_{b_i}(t)$, as the amount of workload sent to a data center to be served by brown energy. Also, the amount of queue length that has to be served using brown energy, $\delta_{b_i}(t)$, is specified as $\delta_{b_i}(t) = \delta_i(t) - \delta_{g_i}(t)$.

When using brown energy, we note the different deregulated electricity markets of data centers located at different regions. Denote $C_i(t)$ as the price of electricity at data center i at time t. In order to benefit from electricity price diversity, the distribution center can employ the day-ahead electricity price forecasting methods [9], [10].

III. Optimization Framework

A. Green Workload Distribution

In this section, we propose an optimization framework to facilitate green workload distribution by using our definition of green service rate, green workload and $\delta_{g_i}(t)$. Our framework employs the results of the power generation forecasting methods as the input. The objective of green workload distribution is to maximize the utilization of green energy at each time slot.

The idea of our optimization is to distribute the workload to data centers based on the residual green service rate, i.e., the green rate which is not used to serve requests in the queues. That is, in order to maximize the utilization of green energy, the following problem is proposed to be solved at the workload distribution center at the beginning of each time slot (e.g., every few minutes),

$$\underset{\lambda_{g_i}(t)}{\text{minimize}} \quad \sum_{i=1}^{N}(\lambda_{g_i}(t) - \mu_{g_i}(t) + \delta_{g_i}(t))^2 \tag{2}$$

$$\text{subject to} \quad \sum_{i=1}^{N}\lambda_{g_i}(t) \leq L(t), \tag{3}$$

$$0 \leq \lambda_{g_i}(t) \leq \mu_{g_i}(t) - \delta_{g_i}(t), \ \forall i = 1, \ldots, N. \tag{4}$$

This problem is a linear least square problem with inequality constraints and different existing algorithms can be used to find the solution. While the objective is to maximize the green energy utilization via optimizing $\lambda_{g_i}(t)$, the constraints (3) and (4) are to limit the allocated green workload to the data centers by the total incoming workload and available green resources, respectively. Thus, the optimal solution at each time slot, i.e., $\overset{\star}{\Lambda}_G(t) = [\overset{\star}{\lambda}_{g_1}(t), \ldots, \overset{\star}{\lambda}_{g_N}(t)]$, is allocated green workloads to data centers.

B. Brown Workload Distribution

In this section, we propose an optimization framework for brown workload distribution. The objective of our framework is to minimize the total electricity cost. We apply our optimization in each time slot and consequently a distribution strategy allocates the so-called brown workload to the data centers that minimize the electricity cost. That is, if the green energy is not adequate to serve all incoming workload, we solve the following linear program at the beginning of each time slot,

$$\underset{\lambda_{b_i}(t),\ \mu_{b_i}(t)}{\text{minimize}} \quad \sum_{i=1}^{N} C_i(t)\mu_{b_i}(t) \tag{5}$$

$$\text{subject to} \quad \sum_{i=1}^{N}\lambda_{b_i}(t) = L(t) - \sum_{i=1}^{N}\lambda_{g_i}(t), \tag{6}$$

$$\lambda_{b_i}(t) + \delta_{b_i}(t) - \mu_{b_i}(t) \leq D_i(t), \ \forall i = 1, \ldots, N, \tag{7}$$

$$\mu_{b_i}(t) \geq 0, \ \lambda_{b_i}(t) \geq 0, \ \forall i = 1, \ \ldots, N. \tag{8}$$

While we use equality constraint (6) to allot all the incoming workload to the data centers, the inequality constraints (7) are to enforce QoS requirements at each data center. We set $D_i(t)$ of each data center proportional to its brown workload, i.e., $D_i(t) = q\lambda_{b_i}(t)$. In this way, we maintain an upper bound on our estimation of queue length at the next time slot, i.e., $\lambda_{b_i}(t) + \delta_{b_i}(t) - \mu_{b_i}(t)$. The value of q is chosen small enough to enforce the QoS requirements at all data centers. Also, we assume that the total number of servers at each data center, M_i, is large enough. Therefore, $\overset{\star}{\Lambda}_B(t) = [\overset{\star}{\lambda}_{b_1}(t), \ldots, \overset{\star}{\lambda}_{b_N}(t)]$ and $\overset{\star}{M}_b(t) = [\overset{\star}{\mu}_{b_1}(t), \ldots, \overset{\star}{\mu}_{b_N}(t)]$, the solution to this problem, are respectively the assigned brown workload and service rate to the data centers.

IV. Workload Distribution Algorithm

In this section, based on the proposed optimization problems, we design an algorithm to allocate workload to data centers in a green and low-cost manner. As depicted in Algorithm 1, we first offer the green workload distribution problem to compute optimum green workload allocated to each data center. In the case that green energy is not enough to serve all incoming workload, we solve the brown workload distribution problem to apportion more workload to each data center to be served by brown energy. Accordingly, the output of the proposed algorithm is the total assigned workload to each data center, i.e., green and brown workload.

Algorithm 1 GOOD Algorithm

INPUT: The available green energy at each data center, $W_i(t)$
The price of electricity at each data center, $C_i(t)$
The total incoming workload, $L(t)$
The QoS parameter, q
OUTPUT: The allocated workload to each data center, $\lambda_i(t) = \lambda_{g_i}(t) + \lambda_{b_i}(t)$

```
for each time slot do
    green workload distribution optimization and, λ_{g_i}(t) = λ*_{g_i}(t), ∀i = 1, ..., N
    if Σ_{i=1}^{N} λ_{g_i}(t) < L(t) then
        solve brown workload distribution optimization and, λ_{b_i}(t) = λ*_{b_i}(t), ∀i = 1, ..., N
    else
        λ_{b_i}(t) = 0, ∀ i = 1, ..., N
    end if
    λ_i(t) = λ_{g_i}(t) + λ_{b_i}(t), ∀i = 1, ..., N
end for
```

V. Simulation Results

We consider $N = 3$ data centers, each integrated with a wind farm as a renewable power source. It is assumed that the data centers are located at three different regions with deregulated electricity market. Our simulation data are based on the trends of wind power and electricity price as shown in Figure 2 [2]. We simulated the total workload using a sample day of the requests made to the 1998 World Cup web site [11]. Figure 3 compares the electricity cost of running the three data centers for an instance QoS parameter $q = 0.05$. As depicted in this figure, the electricity cost of the proposed workload distribution algorithm outperforms the uniform workload distribution and the scheme in [2] for the same QoS parameters. Also, the allocated green workloads and the sustained queue lengths at the data centers are shown in Figure 4. Finally, Figure 5 is provided to show the total allocated green and brown workload to each data center. For example, the trend of price of electricity and wind power indicates that before hour 11 most of the workload is assigned to data center 2. However, from hours 11 to 14, the price of electricity at data center 3 is lower than those at the other data centers, and thus most of the workload is allocated to this data center. In addition, for instance, at hour 24, the total

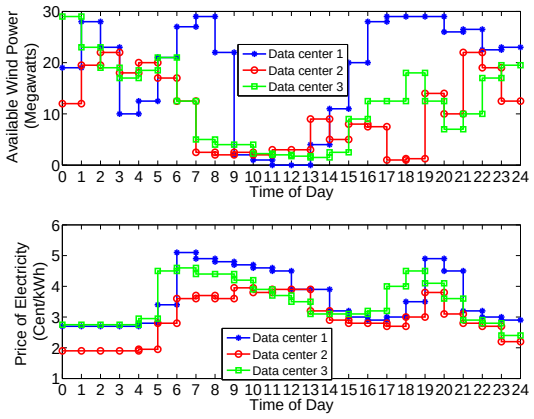

Fig. 2. Wind power generation and price of electricity.

incoming workload is low and it is possible to serve most of the workload as the green workload. In other words, in this hour, the available wind power is the key decision factor to allocate workloads among data centers. Therefore, these changes in workload distribution lead to a reduction in the total cost of running the three data centers.

VI. Conclusion

The separation of green energy utilization maximization and brown energy cost minimization problems is proposed by decomposing the workload into that served by green and brown energy, respectively. In this regard, a new notion of green workload and service rate versus brown workload and service rate for data centers has been introduced, and accordingly a new distribution algorithm has been designed and demonstrated to outperform the existing workload distribution strategies in terms of the cost of electricity.

References

[1] Y. Zhang and N. Ansari, "On architecture design, congestion notification, tcp incast and power consumption in data centers," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 39–64, 2013.

[2] M. Ghamkhari and H. Mohsenian-Rad, "Optimal integration of renewable energy resources in data centers with behind-the-meter renewable generator," in *IEEE International Conference on Communications (ICC)*, 2012, pp. 3340–3344.

[3] M. Ghamkhari and H. Mohsenian-Rad, "Energy and performance management of green data centers: A profit maximization approach," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 1017–1025, 2013.

[4] 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 *Proceedings IEEE INFOCOM*, 2010, pp. 1–9.

[5] J. Li, Z. Li, K. Ren, and X. Liu, "Towards optimal electric demand management for internet data centers," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 183–192, 2012.

[6] United States Environmental Protection Agency, "Epa report on server and data center energy efficiency," Final Report to Congress, Aug. 2007.

[7] S. S. Soman, H. Zareipour, O. Malik, and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," in *IEEE North American Power Symposium (NAPS)*, 2010, pp. 1–8.

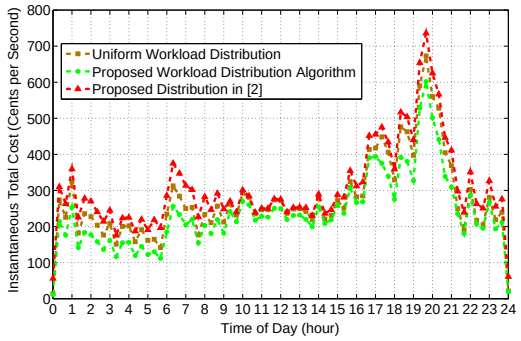

Fig. 3. Instantaneous total cost of electricity for a sample day.

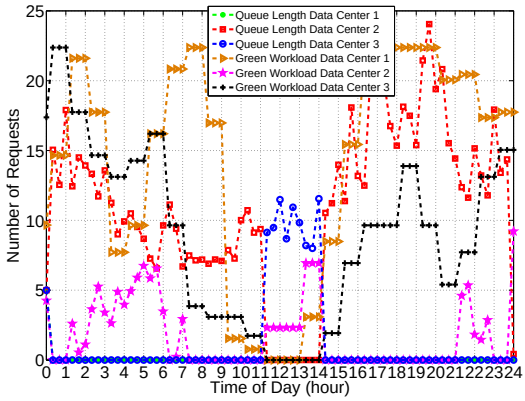

Fig. 4. The allocated green workloads and the sustained queue lengths at the data centers.

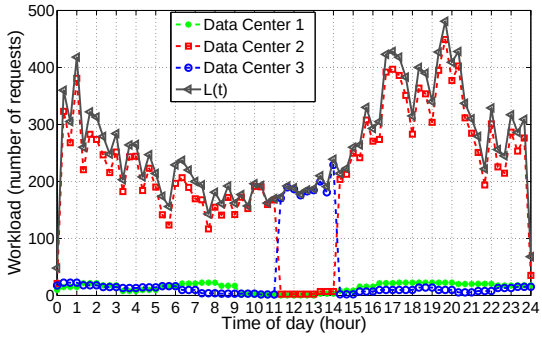

Fig. 5. Allocated workload to the data centers by the proposed workload distribution algorithm.

[8] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2011, pp. 528–533.

[9] L. Wu and M. Shahidehpour, "A hybrid model for day-ahead price forecasting," *IEEE Transactions on Power Systems*, vol. 25, no. 3, pp. 1519–1530, 2010.

[10] P. Areekul, T. Senjyu, H. Toyama, and A. Yona, "A hybrid arima and neural network model for short-term price forecasting in deregulated market," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 524–530, 2010.

[11] http://ita.ee.lbl.gov/html/contrib/WorldCup.html.