Optimizing Green Energy, Cost, and Availability in Distributed Data Centers

Rakesh Tripathi, S. Vignesh, and Venkatesh Tamarapalli

***Abstract*— Integrating renewable energy and ensuring high availability are two major requirements for geo-distributed data centers. Availability is ensured by provisioning spare capacity across the data centers to mask data center failures (either partial or complete). We propose a mixed integer linear programming formulation for capacity planning while minimizing the total cost of ownership (TCO) for highly available, green, distributed data centers. We minimize the cost due to power consumption and server deployment while targeting a minimum usage of green energy. Solving our model shows that capacity provisioning considering green energy integration not only lowers carbon footprint but also reduces the TCO. Results show that up to 40% green energy usage is feasible with a marginal increase in the TCO compared with the other cost-aware models.**

***Index Terms*— Geo-distributed green data center, cloud provisioning, operating cost minimization, fault tolerance.**

I. Introduction

With increased usage of Internet services, there is a rapid growth in the number of geo-distributed data centers around the world. At the same time, data center operators are under pressure to minimize the carbon footprint. One of the ways to do this is to use renewable energy from on-site or off-site sources. Recently, all major distributed data-centers are powered either partially or fully by renewable energy. Gao *et al.* [1] demonstrated how distributed data centers can exploit uncorrelated wind sources to meet 95% of their energy requirement. The work in [2] minimized the cost of building data centers and renewable energy consumption while satisfying constraints on green energy integration, availability and latency. Ren *et al.* [3] considered capping carbon footprint while minimizing the operating cost (including utility price, renewable energy cost, battery cost and operational expenditure). For a good survey of the literature dealing with integration of renewable energy in data centers, see [4].

It is estimated that the financial loss for an hour of downtime can range from $250,000-$500,000 for a large data center operator [5]. Typically, high availability (also termed fault-tolerance) is handled by spare capacity provisioning to mask partial or complete data center failure (at a site). While the work in [6] minimizes the cost of spare capacity provisioning by minimizing the number of servers, our previous work showed the need to minimize the operating cost by considering spatio-temporal variation in electricity price [7]. However, none of them considered the cost of green energy procurement while optimizing the operating cost.

Manuscript received September 19, 2016; revised November 2, 2016; accepted November 8, 2016. Date of publication November 22, 2016; date of current version March 8, 2017. The associate editor coordinating the review of this letter and approving it for publication was O. Amin.

The authors are with the Department of Computer Science & Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India (e-mail: t.rakesh@iitg.ernet.in; s.vignesh@iitg.ernet.in; t.venkat@iitg.ernet.in).

Digital Object Identifier 10.1109/LCOMM.2016.2631466

Existing literature mostly addressed workload distribution considering green data centers to minimize the total cost of ownership (TCO). Our work is the first one to consider the real-time price of electricity (while enforcing a minimum green energy usage), to design cost-efficient fault-tolerant distributed data centers. We provision spare capacity across the data centers so that the demand is met even after the failure at a data center site (either partial or complete), while minimizing the TCO. We consider both green and brown energy cost in minimizing the operational cost. We model the problem using mixed integer linear programming (MILP) where, the main constraints include; green energy usage bound, latency bound, and the failure probability at a site (partial or complete). Solving our model gives the optimal server distribution across the sites and the demand distribution that minimizes the TCO.

II. Optimization Framework

In this section, we first state the assumptions in the model for green energy availability, failure and power consumption, and then present the MILP formulation.

A. Assumptions

We assume a distributed data center powered by multiple green and brown energy sources. Each data center is integrated with green energy sources such as, onsite/offsite renewable energy sources (wind and/or solar), Power Purchase Agreement (PPA), and brown energy through grid. The brown energy and PPA can be used flexibly based on the power demand and available green energy. To account for different types of workload, we consider heterogeneous demand. The following assumptions are used in the model.

- Each data center consolidates the workload to keep the power consumption proportional to the workload served.
- Service time (including queuing delay) is same for all the data centers and the latency is due to the propagation delay.
- The requests are placed in a single queue to be served by any server. All servers are uniformly loaded.
- Only one data center can completely or partially fail at any point in time [6]. Failure of more than one data center at the same time is avoided by the choice of locations.

B. Optimization Problem Formulation

1) Demand: Let S be the set of data centers housing m_s number of servers. Let L_u^{ah} denote the demand from a client region u during hour h for application type a. Let λ_{su}^{afh} denote the number of requests mapped from client region u to data center s ($s \in S$), at hour h for an application type a. Here $f = 0$ indicates the case of no data center failure and $f \in 1, 2..|S|$ indicates the case of f^{th} data center failure.

2) Heterogeneous Workload: Since we assume that a data center can serve different types of workload, we explicitly considered the heterogeneity while calculating the server utilization. Let the processing rate of the server be B bps and the mean job size be J_a for application type a. The effective service rate is $\frac{B}{J_a}$. We define the average utilization as

$$\gamma_s^{fh} = \frac{\sum_{u,a} \lambda_{su}^{afh} J_a}{m_s B} \tag{1}$$

3) Failure Model: Let p be the fraction of servers failing at any given site. The processing rate of a failed data center reduces to $(1-p)m_s B$. The data center utilization after the failure can be expressed as

$$\gamma_s^{fh} = \begin{cases} Eq.(1) & \forall f \neq s \\ \dfrac{\sum_{u,a} \lambda_{su}^{afh} J_a}{(1-p)m_s B} & f = s,\ p < 1 \\ 0 & f = s,\ p = 1 \end{cases} \tag{2}$$

4) Delay: Let D_{su} be the propagation delay between the data center s and the client region u. We define a target delay of D_{max} for all types of workloads (with different processing rate), when no data center has failed and D_{max}^f ($D_{max}^f \geq D_{max}$) for the case of a data center failure. We use a binary variable, y_{su}^f to indicate the ability of data center s to service requests from client region u when data center f has failed.

5) Power Consumption: Let P_{idle} be the average power drawn in idle condition and P_{peak} be the power consumed when server is running at peak utilization. The total power consumed by $s \in S$, at hour $h \in H$ is modeled as [8]

$$\begin{aligned} P_s^{fh} &= m_s(P_{idle} + (E_s - 1)P_{peak}) \\ &\quad + m_s(P_{peak} - P_{idle})\gamma_s^{fh} + \epsilon, \end{aligned} \tag{3}$$

where E_s is the power usage effectiveness (PUE) of a data center at s, defined as the ratio of the total power consumed by the data center to the power consumed by the computing equipment, and ϵ is an empirical constant.

6) Modeling Brown Energy Usage: Let θ_s^h be the price of brown energy at a data center s during hour h and δ_{si}^h be the price of green energy of type i, $i \in \{1,2,3,4,5\}$ corresponding to onsite wind, offsite wind, onsite solar, offsite solar and PPA, respectively. Let PB_s^{fh} denote the amount of brown energy drawn at hour h and Δ_{si}^{fh} denotes the amount of renewable energy drawn from source i. Since the brown energy is used only after exhausting the green energy available, the brown energy drawn from the grid is given by

$$PB_s^{fh} = P_s^{fh} - \sum_i \Delta_{si}^{fh} \quad \forall s, h, f \tag{4}$$

7) Cost Model: We first define the following cost components.

- *Server cost:* Let α be the cost of acquiring a server. The total cost of the servers in all the data centers is

$$\Phi = \alpha \sum_s m_s \tag{5}$$

- *Brown energy cost:* The cost of brown energy consumed across all the data centers is given by

$$\Theta = \sum_{s,h,f} \theta_s^h PB_s^{fh} \tag{6}$$

- *Renewable energy cost:* The total cost incurred in using renewable energy across all the data centers is given by

$$R = \sum_{s,h,f} \delta_{si}^h \Delta_{si}^{fh} \tag{7}$$

8) Objective Function: The objective for capacity provisioning in fault-tolerant green distributed data centers is to minimize the TCO, denoted by F, which is simply the sum of all the aforementioned costs while satisfying constraints on delay, green energy usage and availability. Formally, the problem is expressed as

$$\textbf{minimize } F = \Phi + \Theta + R; \tag{8}$$

subject to

$$\sum_{s,i,h} \Delta_{si}^{fh} \geq \rho P_s^{fh}, \quad \forall f \tag{9}$$

$$\sum_s \lambda_{su}^{afh} = L_u^{ah}, \quad \forall u, a, h, f \tag{10}$$

$$2D_{su}\, y_{su}^f \leq D_{max}, \quad \forall s, u, f = 0 \tag{11}$$

$$2D_{su}\, y_{su}^f \leq D_{max}^f, \quad \forall s, u, f \geq 1 \tag{12}$$

$$0 \leq \lambda_{su}^{afh} \leq y_{su}^f L_u^{ah}, \quad \forall s, u, a, h, f \tag{13}$$

$$\gamma_s^{fh} \leq \gamma^{max}, \quad \forall s, h, f \tag{14}$$

$$M^{min} \leq m_s \leq M^{max}, \quad \forall s \tag{15}$$

$$\lambda_{su}^{afh} = 0, \quad \forall u, a, h, s = f \tag{16}$$

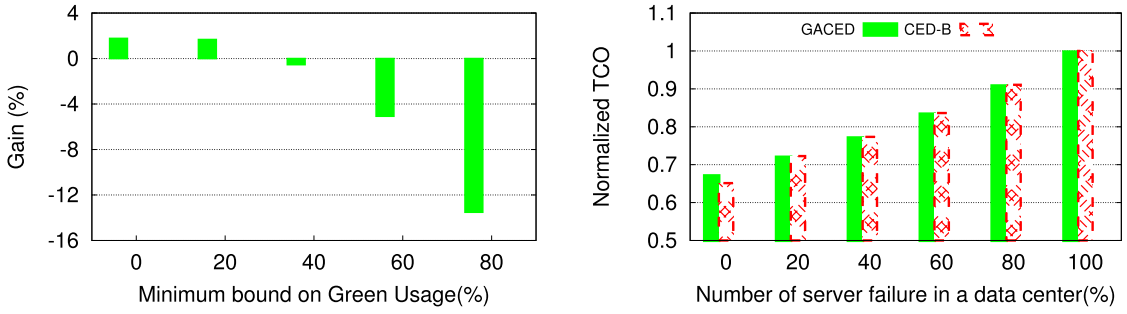
$$y_{su}^f \in \{0,1\}, \quad \forall s, u, f \tag{17}$$

Among the constraints, Eq. 9 ensures that the green energy sources meet ρ percent of the total power demand over the time. Eq. 10 makes sure that the demand during every hour is met. Eq. 11,12, and 13 ensure that all types of workload are served within the latency bound (before and after failure). Eq. 14 is used to limit the queuing delay by bounding the average server utilization to $\gamma^{max} \in (0,1]$. It also ensures that workload assigned to a failed data center is bounded by its capacity. Eq. 15 ensures that capacity limit of a data center (in terms of number of servers) is not exceeded. Eq. 16 ensures that no request is served by a failed data center.

The decision variables in the MILP are: m_s, the number of servers in a data center s, λ_{su}^{afh}, the number of requests from client region u mapped to data center s at hour h for workload type a, and Δ_{si}^{fh}, the renewable energy drawn from source i.

III. Numerical Results

The proposed MILP model (termed GACED) is solved centrally using CPLEX and MATLAB tools on a Linux server with Intel Xeon processor and 64 GB of RAM. Since spare capacity provisioning in data centers is a one-time effort at the time of design, the running time is not a matter of concern. We obtained the necessary data for three locations: Texas, Illinois, and California from the Web. The server processing rate is set to 1.6 MBps. Two types of workload are considered



(a) Green usage bound (b) Failure percentage

Fig. 1. Impact of varying green energy usage and failure percentage on the TCO of GACED model.

TABLE I
CAPACITY FACTOR FOR VARIOUS GREEN ENERGY SOURCES

	Source	Location	Avg CF(%)
Onsite	Wind	California	27
	Wind	Illinois	32
	Wind	Texas	33.6
	Solar	California	22.8
	Solar	Texas	23.46
Offsite	Wind	Arizona	30
	Wind	Colorado	43
	Wind	Iowa	34
	Solar	Arizona	27.74
	Solar	Colorado	24.61

with the mean job size 13KB and 26KB, and the service rate of 120 and 60 requests/sec. D_{max} and D^f_{max} are set to 40ms and 80ms, respectively. PPA price for TX, CA, IL is taken as 8, 8, 4 cents/kWh, respectively. For other parameters refer [7]. The client demand was generated from traces of Wikipedia.org.[1] We chose nine regions of USA: Illinois, Tennessee, New York, Arizona, Massachusetts, California, Florida, Missouri, and Louisiana. The demand at each location was proportional to the number of Internet users.[2] Table I reports on-site and off-site renewable energy sources at each location with the corresponding average capacity factor (CF), defined as the ratio of the actual power output to the maximum rated capacity.

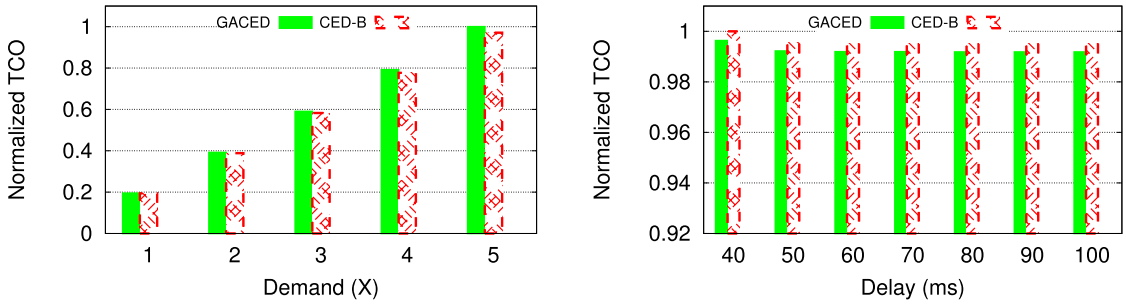
We used the models from NREL[3] and [9] for solar and wind energy generation, respectively. Based on the meteorological data from NREL,[4] we calculated the total power generated. At each site, we considered 20 wind turbines of capacity 1.5MW, each and 10,000 solar panels of 120W, each. The cost of generating wind and solar power is obtained by taking the installation cost of 1630 and 3100 $/kW, and life time of 20 and 25 yrs, respectively [3]. We took quarterly average of client demand and renewable energy generated for every hour of the day. The brown electricity price at different locations is taken from US energy information administration website.

For comparison, we designed a baseline model (termed CED-B) that minimizes the TCO, where the data centers are powered only with brown energy while retaining other constraints. For a full-site failure scenario, Fig. 1a shows the percentage gain in the TCO using GACED model compared to the CED-B model. Even after failure, the gain is 2% with the green energy usage of 20%. The gain reduces with increase in green energy usage because, our model increases the amount of (expensive) green energy purchased to satisfy the constraint. On the other hand, CED-B model has no cost from green energy usage. With the GACED model greening can be achieved with very little or no extra cost, unless we target high renewable energy usage.

Fig. 1b compares the TCO for the two models while forcing 40% green energy usage and varying the failure percentage. We see that the TCO is almost similar for both the models due to the fact that, the GACED model optimally uses cheaper renewable energy to reduce the TCO. Fig. 3 illustrates the offsite wind energy usage in the GACED model and its corresponding price at the Texas data center. When the wind energy is cheaper, GACED uses more of it to maintain the same TCO (as with CED-B), albeit with reduced carbon footprint. Due to intelligent usage of green energy, it was possible to meet the target renewable energy usage of 40% at all times. We conclude that the GACED model can lead to greener data center deployment with no or little additional cost (though green energy procurement is costlier).

Fig. 2a shows the impact of increase in demand (multiples of baseline demand in previous experiment) on the TCO and green energy procurement decision. We set green energy usage and failure percentage to 40% and 20%, respectively. As demand increases, the TCO for both the models increases due to obvious reasons. However, TCO with GACED model increases with the demand due to the green energy usage constraint. We note that, even with five-fold increase in the demand (with a cost of almost 30 cents/kW for wind energy at Texas), it is possible to meet the 40% green energy usage constraint with a meagre 4% increase in the TCO. Fig. 2b shows the impact of relaxing latency requirement on the TCO. D^f_{max} is set to twice D_{max}. We notice that both the models reduce the TCO for relaxed latency bound, since there is more choice in the data centers serving a region. However, GACED model lowers the TCO by considering locations powered by cheaper green energy.

[1] http://dumps.wikimedia.org/other/pagecounts-raw
[2] http://www.internetworldstats.com/unitedstates.htm
[3] http://pvwatts.nrel.gov/pvwatts.php
[4] http://www.nrel.gov/midc/



(a) TCO with demand variation

(b) TCO with delay bound

Fig. 2. Impact of demand and delay bound variation on the TCO of GACED model.

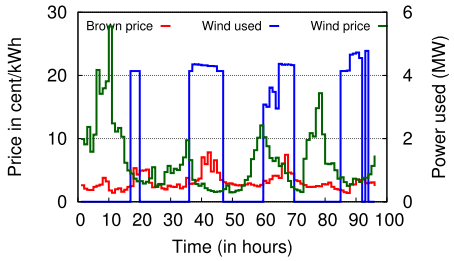


Fig. 3. Illustration of wind energy usage.

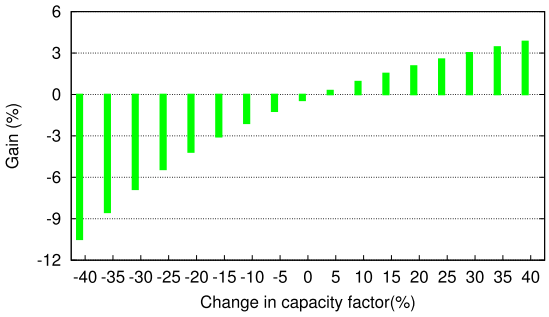


Fig. 4. Gain in TCO for GACED and CED-B models varying CF.

A. Sensitivity Analysis

We quantitatively evaluate the impact of uncertainty in the renewable energy availability on the performance of GACED model. For the case of complete data center failure and 40 % green energy usage requirement, the capacity factor was varied in the range of [-40%, 40%]. Fig. 4 shows the percentage gain in the TCO with GACED model (compared to the CED-B model). Since, the cost of renewable energy decreases with increasing capacity factor [3], the GACED model has lower TCO compared to the CED-B model (by about 4%). This is because it efficiently exploits cheaper green energy. If the forecasted green energy availability is inaccurate, *i.e.*, the capacity factor changes by -40%, we can work with 20% green energy, where GACED model has only 3% higher TCO.

Harnessing green energy depends upon the cost and efficiency of technology, which is constantly improving thereby reducing the cost. For example, between 2006 and 2014, the world-wide average photo-voltaic solar cell cost had dropped by 78%.[5] Considering about 10% reduction in green energy price per year, we computed that the GACED model leads to a gain of 10% in the TCO even with 80% green energy usage.

[5]http://solarcellcentral.com/cost_page.html

IV. Conclusion

We used MILP to formulate the cost-aware capacity provisioning problem for fault-tolerant data centers ensuring a minimum green energy usage (GACED model). The proposed model outperforms the baseline model (CED-B) which minimizes the TCO considering only brown energy. Results demonstrate that even with renewable energy integration, the TCO can be low with GACED model, despite green energy being costlier. Even forcing a green energy usage of upto 80% leads to an additional cost of only 15% compared to the CED-B model. The proposed model optimally schedules demand considering the availability of green energy and its price variation to lower the TCO. We conclude that with an appropriate model, green energy integration lowers the cost of designing fault tolerant distributed data centers with reduced carbon footprint. Our model would be further beneficial with an improvement in technology leading to larger capacity factor (lower renewable energy cost).

References

[1] V. Gao, Z. Zeng, X. Liu, and P. R. Kumar, "The answer is blowing in the wind: Analysis of powering Internet data centers with wind energy," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 520–524.

[2] J. L. Berral, Í. Goiri, T. D. Nguyen, R. Gavaldá, J. Torres, and R. Bianchini, "Building green cloud services at low cost," in *Proc. IEEE ICDCS*, Jun. 2014, pp. 449–460.

[3] C. Ren, D. Wang, B. Urgaonkar, and A. Sivasubramaniam, "Carbon-aware energy capacity planning for datacenters," in *Proc. IEEE MASCOTS*, Aug. 2012, pp. 391–400.

[4] W. Deng, F. Liu, H. Jin, B. Li, and D. Li, "Harnessing renewable energy in cloud datacenters: Opportunities and challenges," *IEEE Netw.*, vol. 28, no. 1, pp. 48–55, Jan./Feb. 2014.

[5] "2013 study on data center outages." [Online]. Available: http://www.emersonnetworkpower.com/documentation/en-us/brands/liebert/documents/whitepapers/2013_emerson_data_center_outages_sl-24679.pdf

[6] I. Narayanan, A. Kansal, A. Sivasubramaniam, B. Urgaonkar, and S. Govindan, "Towards a leaner geo-distributed cloud infrastructure," in *Proc. HotCloud*, 2014, pp. 1–8.

[7] R. Tripathi, S. Vignesh, and V. Tamarapalli, "Cost-aware capacity provisioning for fault-tolerant geo-distributed data centers," in *Proc. COMSNETS*, Jan. 2016, pp. 1–8.

[8] A.-H. Mohsenian-Rad and A. Leon Garcia, "Energy-information transmission tradeoff in green cloud computing," in *Proc. IEEE GLOBECOM*, 2010, pp. 1–6.

[9] Y. Zhang, Y. Wang, and X. Wang, "Greenware: Greening cloud-scale data centers to maximize the use of renewable energy," in *Proc. ACM/USENIX Middleware*, 2011, pp. 143–164.