Renewable Energy Curtailment via Incentivized Inter-datacenter Workload Migration

Ahmed Abada[(✉)] and Marc St-Hilaire

Department of Systems and Computer Engineering,
Carleton University, Ottawa, ON, Canada
{ahmed.abada,marc_st_hilaire}@carleton.ca

Abstract. Continuous Grid balancing is essential for ensuring the reliable operation of modern smart grids. Current smart grid systems lack practical large-scale energy storage capabilities and therefore their supply and demand levels must always be kept equal in order to avoid system instability and failure. Grid balancing has become more relevant in recent years following the increasing desire to integrate more Renewable Energy Sources (RESs) into the generation mix of modern grids. RESs produce intermittent energy supply that can't always be predicted accurately [1] and necessitates that effective balancing mechanisms are put in place to compensate for their supply variability [2,3]. In this work, we propose a new energy curtailment scheme for balancing excess RESs energy using data centers as managed loads. Our scheme uses incentivized inter-datacenter workload migration to increase the computational energy consumption at a destination datacenter by the amount necessary to balance the grid. Incentivised workload migration is achieved by offering discounted energy prices (in the form of Energy Credits) to large-scale cloud clients in order to influence their workload placement algorithms to favor datacenters where the energy credits can be used. Implementations of our system using the CPLEX ILP solver as well as the Best Fit Decreasing (BFD) heuristic [4] for workload placement on data centers showed that using energy credits is an effective mechanism to speed-up/control the energy consumption rates at datacenters especially at low system loads and that they result in increased profits for the cloud clients due to the higher profit margins associated with using the proposed credits.

1 Introduction

The lack of large-scale energy storage solutions in modern smart-grids necessitates that their energy generation and consumption levels are always kept equal in order to avoid system instability and service interruptions. Since making online fine adjustments to the energy levels of the generators supplying the grid can be costly (because of the different start-up/shut-down delay constraints of each generator and the monetary costs associated with making such adjustments),

many Demand-Response (DR) proposals were introduced by the utility companies to exploit the demand-side flexibility of energy consumers for grid balancing. DR programs generally aim to control the consumers demand for electricity by using monetary incentives such as variable pricing or time-of-use pricing [5–7]. Commercial class energy consumers have been the main focus of many DR proposals because of their higher consumption rates and the higher demand elasticity of some commercial consumers compared to their residential counterparts. Datacenter plants in particular have received a significant share of DR proposals because of their consistent rapid growth in recent years (as a result of the increasing popularity of cloud computing applications) and their ability to control their energy consumption levels by using different energy management techniques or adjusting the scheduling of their computational workloads [8–11]

Data centers are known to consume a lot of energy to operate since they typically comprise several thousands of computing and networking equipments in a physically monitored environment [12]. The high energy consumption of datacenters has attracted many optimization research efforts [13]. Traditionally the main focus has been on minimizing their energy consumption cost either by sever consolidation [14–16], using DVFS [17] or by proposing different workload scheduling/migration techniques that take advantage of variable real-time energy prices to process more workloads when the prices are low [18,19]. More recently, following the growing demand to integrate more green RESs into the generation mix of modern smart grids, more attention was given to datacenters as controllable loads that could play a major role in enabling the integration of more RESs by compensating for their intermittent supply and help maintain the grid balance. Such DR related research is mainly concerned with the effective integration of data centers as a "resource" or a "managed load" into the various smart grid based DR programs [5,8,20].

Because of their highly intermittent supply, it can be a challenge for the grid to maintain its balance at times of excess RESs energy generation. Hence, the ability of the grid to achieve timely and effective renewable energy curtailment have been a distinctive hurdle that needed to be addressed more effectively in order to increase the amount of renewable energy deployment into the generation mix of modern grids [21]. Events of unexpected excess RESs energy generation can lead to system instability and utility companies opt at times to apply negative energy prices in order to re-balance the grid in a timely manner. However, grid failures can still occur due to such events [22,23].

The curtailment of renewable energy have been studied in the past to ensure balanced grid operation [3,24]. However, unlike existing curtailment strategies that blocks excess RESs energy from entering into the system [25], in this work, we try to leverage the demand side flexibility of datacenters in order to effectively balance the excess RESs energy. Hence, our proposal has the advantage of putting the excess RESs energy to use rather than simply discarding it. We present a new energy curtailment approach that uses data centers as managed loads to allow the grid system more "downward flexibility" (i.e. the ability to cope with extra supply or low demand). Our approach is based on offering dis-

counted energy cost (a monetary incentive in the form of Energy Credits) to large-scale cloud users to incentivize them to migrate their computational workloads to datacenters connected to the part of the grid having excess RESs energy supply. The amount of Energy credits offered is made equal to the amount of energy that a grid would need to consume in a timely manner in order to stay balanced. We formulate our problem as a linear integer programming problem with the objective of minimizing the cost of assigning computational workloads on available datacenters. We present our simulation results from implementing our model using the CPLEX ILP solver as well as the Best Fit Decrteasing (BFD) workload assignment heuristic [4] that can be more practical in large scale operations.

The remainder of this paper is organized as follows: in the next section we briefly discuss previous work related to earlier data center based demand-response efforts and discuss the importance of these efforts as an enabler for integrating more renewable energy sources into the generation mix of modern smart grids. Section 3 introduces our cloud-broker based grid balancing proposal and presents its system model and optimization formulation. We present the performance analysis conducted on our system in Sect. 4 and finally conclude the paper in Sect. 5.

2 Related Work

Since data centers represent a significant load for the smart grids that they draw power from, many research efforts have proposed closer cooperation between the two in order to ensure a smooth operation on both sides [11,26]. An example of such tight coupling between the power grid and the data center is in [27] which tries to save on the datacenter energy cost by deploying backup batteries locally at the datacenter and charging them at times of low electricity prices then using them to power the datacenter when the prices are higher. The significance of data centers as energy loads and their ability to increase/decrease their energy consumption by adjusting their computational workload scheduling represents a valuable opportunity for the smart grid operators to essentially integrate them as managed loads for grid balancing purposes using DR [6,10].

A successful integration between the datacenters and smart grids would allow the datacenters to be considered as a "resource" by the smart grid operator which in turn can then use it for grid balancing purposes. However, a main obstacle towards this integration is that modern data centers are increasingly becoming colocation based [28], where multiple organizations host their computing/networking equipment at a shared physical location in order to save on management and maintenance costs. In such environments, the colocation operator has no direct control on the workload scheduling or the energy-saving/server-consolodation techniques applied by the tenants. This makes the involvement of such data centers in modern DR programs more challenging compared to the traditional case of owner-operated data centers as described in [29]. Collocation operators suffer from the "split-incentive" phenomenon that hinders their participation in DR programs. On one hand, they desire to get the monetary gains

associated with participating in DR programs, yet, they have no direct way to control the energy consumption of their tenants.

Effective grid balancing is also necessary for enabling large-scale integration of RESs into the smart-grid. The integration of such sources into the generation mix of modern smart grids have made significant progress in recent years. However, because of the indeterministic nature of the power generated by such sources (wind/solar) and the unavailability of large-scale energy storage facilities in modern grid systems, their rate of adoption into the generation-mix of modern grids has been mainly limited by the grid's ability to stay in balance given the nature of their intermittent supply [9]. Recent research [1] points to the fact that excess energy from RESs could often need to be discarded in order to keep the grid in balance. In some cases negative pricing is applied just to achieve a timely balancing of the grid and avoid service interruption. Negative/free energy pricing is applied at times of excess RESs supply when it is cheaper to offer the excess energy for free or at a negative price than to shut down generation facilities and risk future service interruptions. The rapid increase of datacenter deployments in recent years is positioned to play an important role in enabling the integration of more RESs into the smart-grids by acting as a managed load that can effectively absorb the introduced supply variability of RESs sources.

A main challenge to overcome in this regard is how to effectively influence the workload scheduling on the datacenter servers in order to comply with the DR requirements. This is not easy to accomplish in current colocation based datacenters since they lack coordination between the colocation operator and the tenants in charge of workloads scheduling on the servers. On one hand, the colocation operator is interested in complying with DR programs so it can qualify for financial incentives, but on the other hand, the colocation tenants are only interested in achieving maximum performance for their workloads since they have no direct relationship with the grid operator and would not receive any financial gain for adjusting their workload if they comply with the requirements of the DR programs. We suggest in this work that establishing a direct relationship between large scale colocation tenants (such as cloud brokers) and the utility company/grid can lead to a more effective DR implementation in the datacenter domain. Previous work in this area has investigated several mechanisms for extending demand response programs to colocation based datacenters [29–31]. However with the continuing demand for increased renewable energy integration in modern smart grids, a tighter coupling between the grid and datacenters (including their tenants) is of great value in order to take full advantage of the datacenters as managed loads.

3   Cloud-Broker Based Grid Balancing

As opposed to existing demand-response approaches that only target data center operators (that don't necessarily have direct control on workload scheduling in the case of colocations), we propose a new mechanism that would allow extending such programs to the enterprise-type tenants (such as Cloud Brokers

and SaaS/PaaS operators) of colocation data centers. Such tenants typically host their hardware at multiple colocations in different geographical locations to achieve high redundancy and shorter response times. Their workloads account for a large share of the total workloads handled by modern colocation data centers. Hence, they can be considered as a resource or a managed load that smart grids can use to scale up/down the energy consumption in order to meet DR goals.

Our approach is based on introducing a "Market" entity as a communication medium between the grid/utility company and the cloud-broker/colocation-tenant as shown in Fig. 1. When a grid/utility company needs to balance a certain amount of excess RESs energy, it generates a number of energy credits equal to the amount of excess RESs energy that needs to be balanced/consumed. Such energy credits are then offered (at a discounted cost) to the cloud brokers/collocation tenants to incentivize them to migrate their workloads from other datacenter locations to a certain datacenter whose energy consumption needs to be increased to balance the excess RESs energy. The proposed energy credits are made available to the brokers via the central market entity ("EC Market") that handles all the energy credits assignments/transactions between the different utility companies and cloud brokers/colocation tenants. Our system model shown in Fig. 1 comprises four main entities, cloud brokers/colocation tenants, colocation data centers, utility companies (grid) and the EC market. We consider a single cloud broker in our model for simplicity, however, the system can be expanded to handle multiple brokers via game theoretic approaches. Each data center is assumed to adopt a different pricing model based on its popularity and the offered performance guarantees. Cloud brokers continuously receive computing workloads (i.e. Requests) of different weights from their clients and they need to optimally assign/schedule the received workloads on the available data centers to maximizes their return.

We assume that each incoming workload (WL) received by a broker has three main attributes, a Computing weight (WL^{cpu}) that describes how much computing resources it requires, a Memory weight (WL^{memory}) that describes how much storage capacity it requires and an Energy Consumption weight (WL^{kWh}) that describes how much energy it is rated to consume per unit time. We further assume that the cloud broker implements a fixed pricing model whereby its clients are charged solely based on the computing and storage weights of their submitted workloads at the rates C_B^{cpu} and C_B^{memory} respectively. Accordingly, the broker's profit P that the broker charges for handling a request j with a computing weight of WL_j^{cpu} and a memory storage weight of WLj^{memory} can be computed as P_j in Eq. (1) per each allocation interval for the time of its execution.

$$P_j = WL_j^{cpu} * C_B^{cpu} + WL_j^{mem} * C_B^{mem} \qquad (1)$$

Datacenters on the other hand can charge the requests/workloads submitted to them via cloud brokers according to the individual pricing model of each datacenter. Additionally, data centers also charge the cloud brokers for the energy

consumption of their submitted workloads. Accordingly, if a certain workload consumes WL_j^{kWh} units of energy to run on a certain data center, the data center would charge the broker that submitted the workload a total of C_j^k as per Eq. (2):

$$C_j^k = WL_j^{cpu} * C_{DC}^{cpu} + WL_j^{mem} * C_{DC_k}^{mem} + WL_j^{kWh} * C_{DC_k}^{kWh} \qquad (2)$$

Our proposal introduces the concept of "energy credits" which represent a certain amount of energy refund that a cloud broker can use to help offset the otherwise regular energy cost (the last term ($WL^{KWh} * C_{DC}^{kWh}$) in Eq. (2)). As a result, using such credits results in reducing the total amount payable by the brokers to the datacenter where the credits are available. Energy credits are issued by the grid operators when they needs to consume a certain amount of excess RESs energy in order to maintain balance. Energy credits can be claimed against the energy consumption charges on the datacenter connected to the smart grid that issued the credits.
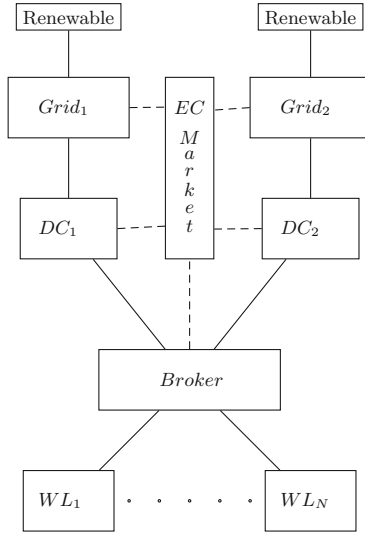
Fig. 1. System model

This direct interaction between the smart grid operators and large scale cloud brokers allows the grid operators to directly influence the scheduling activities of the cloud brokers (such as workload migration from one datacenter to another or workload rescheduling on the same datacenter) on the datacenters that they connect to (by offering reduced energy costs at certain datacenters) and in turn achieve the desired DR objectives. We next introduce the optimization variables and equations of our system model before describing its operation in more detail.

C_B^{cpu} the cost that broker B charges for the processing capacity allocated per time slot (in \$/FLOPS)

C_B^{mem} the cost that broker B charges for disk-space usage per time slot (in \$/MB)

A_{jk}^{next} a binary variable that is set to 1 if job j is assigned to data center k in the next time slot

$A_{jk}^{current}$ a binary variable that is set to 1 if job j is assigned to data center k in the current time slot (initial value = 0)

UEC_{jk} a binary variable that indicates if job j can claim energy credits available at data center k

M the number of available data centers

$C_{DC_k}^{cpu}$ the processing cost at datacenter k per time slot allocated (in \$/FLOPS)

$C_{DC_k}^{mem}$ the disk-space cost at datacenter k per time slot allocated (in \$/MB)

$C_{DC_k}^{kWh}$ the price of energy at datacenter k (in \$/kWh)

EC_k the number of energy credits available at data center k (in kWh)

$MigCost_{kl}$ a normalized value in the range of [0,1] that represents the cost of migrating data from datacenter k to datacenter l (distance-based)

$Cap_{DC_k}^{cpu}$ the maximum processing capacity of datacenter k (in FLOPS)

$Cap_{DC_k}^{mem}$ the maximum disk-storage capacity at datacenter k (in MB)

N the number of client workloads that needs to be assigned to data centers

WL_j^{cpu} the processing capacity needed to process the workload of job j (in FLOPS)

WL_j^{mem} the amount of disk storage needed to host the workload of job j (in MB)

WL_j^{kWh} the energy consumption of job j per each time slot allocated to it (in kWh)

L_{jk} a normalized value (between 0.8–1) that represents the latency between request j and datacenter k. (distance-based)

$$max \ \sum_{k=1}^{M} \left(\sum_{j=1}^{N} \left(\left[WL_j^{mem} * C_B^{mem} + WL_j^{cpu} * C_B^{cpu} \right] * L_{jk} * A_{jk}^{next} \right) \right) \tag{3}$$

$$- \sum_{k=1}^{M} \left(\sum_{j=1}^{N} \left(\left[WL_j^{mem} * C_{DC_k}^{mem} + WL_j^{cpu} * C_{DC_k}^{cpu} + WL_j^{KWh} * C_{DC_k}^{kWH} \right] * A_{jk}^{next} \right) \right) \tag{4}$$

$$+ \sum_{k=1}^{M} \left(\sum_{j=1}^{N} \left(UEC_{jk} * WL_j^{KWH} * C_{DC_k}^{kWH} \right) \right) \tag{5}$$

$$- \sum_{k=1}^{M} \left(\sum_{j=1}^{N} \left(UEC_{jk} * WL_j^{KWH} * 0.1 \ C_{DC_k}^{kWH} \right) \right) \tag{6}$$

$$- \sum_{j=1}^{N} \left(\sum_{k=1}^{M} A_{jk}^{current} \left(\sum_{l=1}^{M} A_{jl}^{next} * MigCost[k][l] * WL_j^{mem} \right) \right) \tag{7}$$

$$s.t. \sum_{j=1}^{N} \left(WL_j^{KWh} * UEC_{jk} \right) < EC_k, \quad \forall k \in M \tag{8}$$

$$\sum_{k=1}^{M} A_{jk}^{next} \leq 1, \qquad\qquad \forall j \in N \tag{9}$$

$$UEC_{jk} <= A_{jk}^{next}, \qquad\qquad \forall j \in N, \; k \in M \tag{10}$$

$$\sum_{j=1}^{N} WL_j^{cpu} * A_{jk}^{next} \leq Cap_k^{cpu}, \qquad \forall k \in M \tag{11}$$

$$\sum_{j=1}^{N} WL_j^{mem} * A_{jk}^{next} \leq Cap_k^{mem}, \qquad \forall k \in M \tag{12}$$

We assume that the cloud broker continuously receives end-user requests (workloads) of different processing (WL^{cpu}), storage (WL^{mem}) and energy consumption (WL^{kWh}) weights and that it needs to find the most cost efficient allocation for the received workloads on the available data centers in order to maximize its revenue. We also assume that each workload and data center in the system is associated with a location coordinate (x, y) that is used to determine the distance-based latency L_{jk} between a workload j and a data center k as well as the migration cost factor $MigCost[k][l]$ between data centers k and l. We consider that time is divided into discrete time intervals of equal duration and that the cloud broker needs to decide a cost efficient allocation for its workloads at the beginning of each interval. We further assume that the pricing coefficients of the cloud broker (C_B^{cpu} and C_B^{mem}) remain fixed over all allocation intervals while the pricing coefficients of the different data centers (C_{DC}^{cpu}, C_{DC}^{cpu} and C_{DC}^{cpu}) do change (within a certain range) from one interval to another in order to simulate a dynamic datacenter pricing and allow the broker to adjust its workload allocation on each interval according to the new prices in order to maximize its revenue. We also assume that the cost of purchasing energy credits can be any fraction of the regular price as dictated by the EC market and consider several example values for this fraction (0.1, 0.4 and 0.7) in our results section to show the effect of this parameter on our model. This allows the broker to have enough incentive to reschedule/migrate its workloads in order to take advantage of the low cost energy and balance the grid in the process.

Our optimization formulation introduced above seeks to maximize the financial gain of the cloud broker subject to system constraints. This optimization is evaluated by the cloud broker at the beginning of each allocation interval in order to decide which data center each of its workloads should be assigned to and whether the allocated workload can benefit from using energy credits or not.

The decision variables of our optimization model are the matrices A_{jk}^{next} and UEC_{jk} introduced earlier. A_{jk}^{next} represents the allocation/assignment result of the broker's workloads (incoming requests) on the different available data centers, whereas UEC_{jk} on the other hand represents whether each workload is counted towards the consumption of energy credits at the datacenters that the

workloads were assigned to (when such credits are available). Our optimization formulation seeks to choose the appropriate values for A_{jk}^{next} and UEC_{jk} such that the objective function is maximized. Assuming that the cloud broker needs to assign a total of N workload requests, each to one of the M available data centers, the above decision variables A_{jk}^{next} and UEC_{jk} can then be represented as two 2-dimensional boolean matrices of size $[N][M]$ where each element A_{jk} (s.t. $j \in N$ and $k \in M$) is assigned the value of "1" if workload j was assigned to be hosted on data center k in the next allocation interval and "0" otherwise. Similarly, each element UEC_{jk} is assigned the value of "1" if workload j is counted towards the consumption of the energy credits available at data center k in the next allocation interval and "0" otherwise.

Using A_{jk}^{next} and UEC_{jk} as the decision variables in the above optimization formulation, the objective function tries to maximize the financial gain of the cloud broker given the different cost coefficients (C_{DC}^{cpu}, C_{DC}^{mem}, C_{DC}^{kWH}) of the M available data centers and the weights (WL^{cpu}, WL^{mem}, WL^{kWh}) of the N workloads that needs to be assigned. The net financial gain is estimated as the sum of total generated revenues (represented with a "+" sign in the objective function) minus the sum of total costs (represented with a "−" sign in the objective function). The first term (3) in the objective function represents the money that the cloud broker generates from its clients in exchange for hosting their workloads on the different data centers. If a workload WL_j was successfully assigned to be hosted on a data center k (i.e. $A_{jk}^{next} = 1$), that workload is said to be generating revenue in the amount shown by the first term of the objective function. The whole first term is multiplied by a scaling-down factor L_{jk} that ranges between [0.8, 1] and represents the distance-based latency between a workload j and data center k where as the distance between a workload and a data center increases, L_{jk} will decrease to approach (0.8). The purpose of scaling down the first term by L_{jk} is to enforce the cloud broker to favor assigning workloads to data centers that are geographically closer to them in order to minimize the latency. The second term (4) of the objective function represents the money that the cloud broker would need to pay to the data center that the workload was assigned to. It is simply the weights of the workload multiplied by the corresponding cost at the data center and finally multiplied by the allocation decision variable A_{jk}^{next}. The third term (5) of the objective function represents the regular value of the used energy credits if they were to be bought at the regular price. The fourth term (6) is the actual (reduced) cost price paid for acquiring the energy credits. Accordingly, the difference between the third and fourth terms (actual and reduced costs) represents a monetary gain that the cloud broker achieves by using the reduced cost energy credits offered by the smart grid. The last term (7) of the objective function represents the cost of migrating a workload of size WL^{mem} from data center k to data center l. For this term, we utilize the $[M] \times [M]$ size $MigCost$ matrix that contains distance-based migration cost factors between the different data centers. Migration cost factors range from [0, 1] and approach the value of 0 as the distance between data centers gets closer. In order to detect a migration event we use $A_{jk}^{current}$ to

hold the allocation results from the previous iteration and we detect a migration if for a workload WL_j we have $A_{jk}^{current} = 1$ and $A_{jl}^{next} = 1$ and $k \neq l$.

The first optimization constraint (8) states that the total weight of workloads that can claim usage of energy credits on a data center may not exceed the total amount of energy credits available on that data center. The second constraint (9) limits the number of data center assignments that a workload can get to a maximum of "1", since we assume that workloads can only be assigned to a maximum of one datacenter at a time. The third constraint (10) ensures that a workload has to be assigned to a data center in order for it to be counted towards the energy credits consumption at that data center (i.e. energy credits can not be claimed at a certain data center if the workload is not assigned to be hosted there). The fourth and fifth constraints (11), (12) are data center capacity constraints to ensure that each data center does not get assigned more workloads than what its capacity can accommodate.

4   Performance Analysis

Since our request/workload assignment problem is essentially a multidimensional binpacking problem where each request is defined by three dimensions (cpu, memory and kWh), the computational complexity of our problem is known to be NP-Hard [32]. Therefore, we have implemented our system model using two approaches, first, as a linear optimization problem into the CPLEX ILP solver and second, using the more scalable approach of the Best Fit Decreasing (BFD) assignment heuristic, so we can evaluate our model under the two implementations. In this section we present our simulation results that measures two different aspects related to our model, namely the effect of the amount of energy credits introduced in the system on the time needed to balance a certain amount of excess energy and the effect of the amount of introduced energy credits on the revenue generated by a cloud broker. We used a system model consisting of three data centers and a single cloud broker as discussed before to maintain system simplicity. Time is modeled in discrete intervals of equal size and the CPLEX/BFD implementations are ran by the broker at the beginning of each interval to determine the allocation/mapping of the requests/workloads on the different available datacenters. Our shown simulation results are the averages of 100 randomized runs where the cost coefficients (C_{DC}^{mem}, C_{DC}^{cpu}, C_{DC}^{kWH}) and location coordinates of all data centers are randomized (within the specified ranges as shown in Table 1) at the beginning of each allocation interval in order to mimic a dynamic cost system that would require the broker to optimize/adjust its workload allocations at the beginning of each new interval and induce workload migrations between data centers. We have set the range of the datacenter energy consumption cost parameter C_{DC}^{kWH} slightly higher than the other datacenter costs in order to make datacenters with available energy credits more attractive as hosts. On the other hand, the cloud broker adopts fixed cost coefficients (C_B^{mem}, C_B^{cpu}) over all allocation intervals in order to shield its clients from the uncertainty of the volatile live pricing model applied by datacenters.

The fixed costs of the broker are set higher than the randomized datacenter costs to allow it to still make profit under any cost applied by the datacenters. We also show the effect of variable system loads on the measured parameters by using different workload arrival rates (10, 20, 30, 40, 50 new request per allocation interval) for each simulation. The list of the different parameters used in our performance testing are listed in Table 1.

Table 1. Simulation parameters

Parameter	Min	Max
$C_{DC}^{mem}, C_{DC}^{cpu}$	5	15
C_{DC}^{kWH}	25	30
C_{B}^{mem}, C_{B}^{cpu}	100	100
$WL^{mem}, WL^{cpu}, WL^{kwh}$	5	15
$WL^{duration}$	3	5
$Available\ energy\ credits\ (EC)$	0	3000
$Workload\ arrival\ rate$	10	50
$Datacenter\ CPU\ capacity$	5000	5000
$Datacenter\ memory\ capacity$	5000	5000

Figure 2 shows the effect of using different cost fractions (0.1, 0.4 and 0.7) for energy credits prices on the time needed to balance the grid given that a certain amount of energy needs to be consumed. Here we assume that 3000 units of energy needs to be consumed by a datacenter in order to balance the grid and we observe how long it takes (in number of allocation intervals) the datacenter to consume this amount when different cost fractions of the original price are offered as energy credits. We notice that reducing the price of offered energy credits always speeds up the time needed for grid balancing especially under light system loads as the workloads quickly become concentrated at the data center that has the energy credits. We also notice that both CPLEX and BFD perform almost exactly the same in all simulated cases. This is due to the big size difference between the request/workload size requirements and the capacity of datacenters (as shown in Table 1), which is a valid assumption in most realistic situations. Because request sizes are considerably smaller than datacenter capacities and the fact that the BFD algorithm always orders the items in a decreasing fashion before assignment, BFD can achieve as optimal binpacking results as CPLEX except for the last "critical element" that gets rejected for size capacity violations. Therefore, the size of the last element that gets rejected from a bin represents an upper bound on the efficiency of the bin assignment. This upper bound is minimized under the assumption of small request sizes. Figure 3 shows the effect of using different cost fractions (0.1, 0.4 and 0.7) for energy credits prices on the broker's generated revenues. We can see that the generated revenues increase as the cost fraction of energy credits
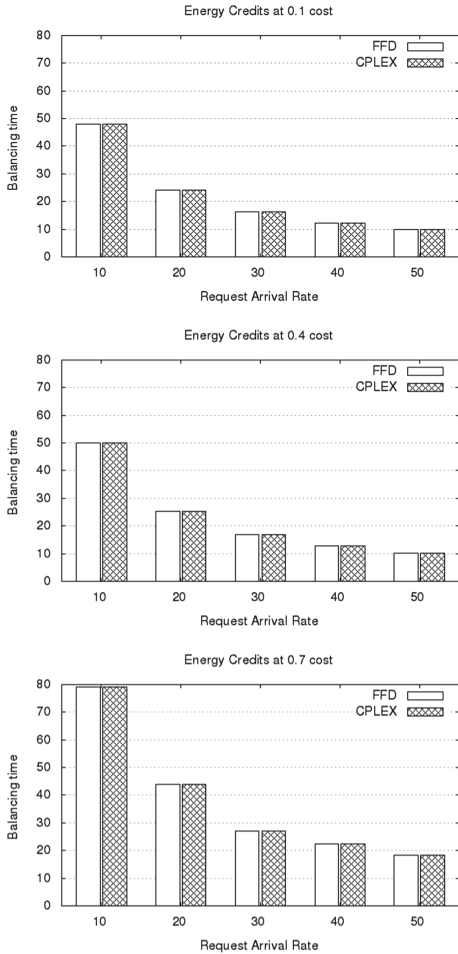
Fig. 2. Effect of varying ECs cost on balancing time
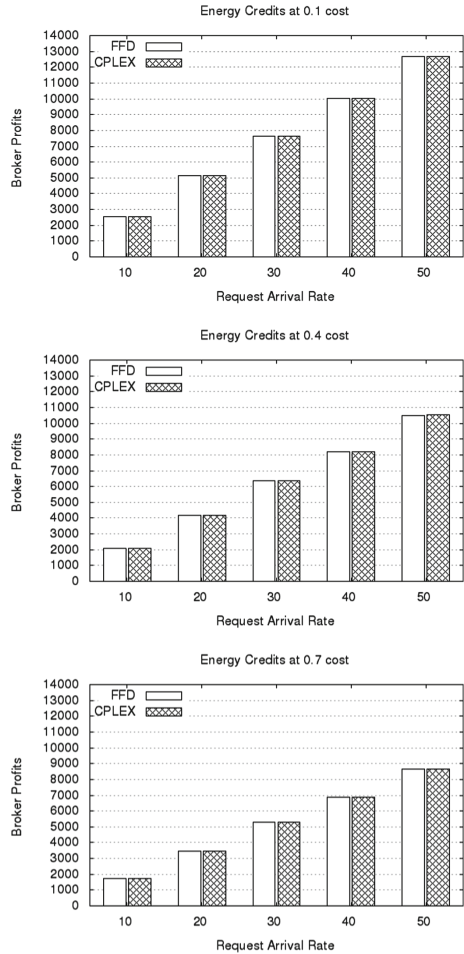
Fig. 3. Effect of varying ECs cost on Broker's profit

decrease. This is expected as energy credits represent discounted energy costs and the lower the cost paid towards acquiring the energy credits the more profits are generated.

5   Conclusion and Future Work

Effective grid balancing is essential for ensuring the reliable operation of modern smart grids. In this work we presented a new linear optimization approach that uses energy credits as an incentive to control the energy consumption levels of large scale datacenter clients (such as cloud brokers/colocation tenant) for grid balancing purposes. Simulations performed on our optimization model using

two different implementation approaches showed that using energy credits as an incentive can speed up the grid balancing process and increases the brokers profit margins. In future work we plan to expand the proposed system to include multiple cloud brokers to compete for the available energy credits by way of game theory bidding mechanism in order to maximize the smart grid's return on the offered energy credits.

References

1. Wang, J., Botterud, A., Bessa, R., Keko, H., Carvalho, L., Issicaba, D., Sumaili, J., Miranda, V.: Wind power forecasting uncertainty and unit commitment. Appl. Energy **88**(11), 4014–4023 (2011)
2. Nicolosi, M.: Wind power integration and power system flexibility-an empirical analysis of extreme events in Germany under the new negative price regime. Energy Policy **38**(11), 7257–7268 (2010). Energy Efficiency Policies and Strategies with regular papers
3. Olson, A., Jones, R.A., Hart, E., Hargreaves, J.: Renewable curtailment as a power system flexibility resource. Electr. J. **27**(9), 49–61 (2014)
4. Khuri, S., Schütz, M., Heitkötter, J.: Evolutionary heuristics for the bin packing problem. In: Pearson, D.W., Steele, N.C., Albrecht, R.F. (eds.) Artificial Neural Nets and Genetic Algorithms, pp. 285–288. Springer, Vienna (1995). https://doi.org/10.1007/978-3-7091-7535-4_75
5. Albadi, M.H., El-Saadany, E.F.: Demand response in electricity markets: an overview. In: 2007 IEEE Power Engineering Society General Meeting, pp. 1–5, June 2007
6. Basmadjian, R., Lovasz, G., Beck, M., Meer, H.D., Hesselbach-Serra, X., Botero, J.F., Klingert, S., Ortega, M.P., Lopez, J.C., Stam, A., Krevelen, R.V., Girolamo, M.D.: A generic architecture for demand response: the ALL4Green approach. In: 2013 International Conference on Cloud and Green Computing, pp. 464–471, September 2013
7. Spees, K., Lave, L.B.: Demand response and electricity market efficiency. Electr. J. **20**(3), 69–85 (2007)
8. Chen, H., Caramanis, M.C., Coskun, A.K.: The data center as a grid load stabilizer. In: 2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC), pp. 105–112, January 2014
9. Chiş, A., Rajasekharan, J., Lundén, J., Koivunen, V.: Demand response for renewable energy integration and load balancing in smart grid communities. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp. 1423–1427, August 2016
10. Kim, H., Kim, Y.J., Yang, K., Thottan, M.: Cloud-based demand response for smart grid: architecture and distributed algorithms. In: 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 398–403, October 2011
11. Wierman, A., Liu, Z., Liu, I., Mohsenian-Rad, H.: Opportunities and challenges for data center demand response. In: International Green Computing Conference, pp. 1–10, November 2014
12. Schomaker, G., Janacek, S., Schlitt, D.: The energy demand of data centers. In: Hilty, L.M., Aebischer, B. (eds.) ICT Innovations for Sustainability. AISC, vol. 310, pp. 113–124. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09228-7_6

13. Rong, H., Zhang, H., Xiao, S., Li, C., Hu, C.: Optimizing energy consumption for data centers. Renew. Sustain. Energy Rev. **58**, 674–691 (2016)
14. Ferreto, T.C., Netto, M.A., Calheiros, R.N., Rose, C.A.D.: Server consolidation with migration control for virtualized data centers. Future Gener. Comput. Syst. **27**(8), 1027–1034 (2011)
15. Varasteh, A., Goudarzi, M.: Server consolidation techniques in virtualized data centers: a survey. IEEE Syst. J. **11**(2), 772–783 (2017)
16. Van, H.N., Tran, F.D., Menaud, J.M.: SLA-aware virtual resource management for cloud infrastructures. In: 2009 Ninth IEEE International Conference on Computer and Information Technology, vol. 1, pp. 357–362, October 2009
17. Wu, C.M., Chang, R.S., Chan, H.Y.: A green energy-efficient scheduling algorithm using the DVFS technique for cloud datacenters. Future Gener. Comput. Syst. **37**, 141–147 (2014)
18. Buchbinder, N., Jain, N., Menache, I.: Online job-migration for reducing the electricity bill in the cloud. In: Domingo-Pascual, J., Manzoni, P., Palazzo, S., Pont, A., Scoglio, C. (eds.) NETWORKING 2011. LNCS, vol. 6640, pp. 172–185. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20757-0_14
19. Rao, L., Liu, X., Xie, L., Liu, W.: Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment. In: 2010 Proceedings IEEE INFOCOM, pp. 1–9, March 2010
20. Chiu, D., Stewart, C., McManus, B.: Electric grid balancing through lowcost workload migration. SIGMETRICS Perform. Eval. Rev. **40**(3), 48–52 (2012)
21. Bollen, M.: The Smart Grid: Adapting the Power System to New Challenges. Morgan and Claypool, San Rafael (2011)
22. Vos, K.D.: Negative wholesale electricity prices in the German, French and Belgian day-ahead, intra-day and real-time markets. Electr. J. **28**(4), 36–50 (2015)
23. Brandstätt, C., Brunekreeft, G., Jahnke, K.: How to deal with negative power price spikes?—flexible voluntary curtailment agreements for large-scale integration of wind. Energy Policy **39**(6), 3732–3740 (2011)
24. Kane, L., Ault, G.: A review and analysis of renewable energy curtailment schemes and principles of access: transitioning towards business as usual. Energy Policy **72**, 67–77 (2014)
25. Subramanian, A., Bitar, E., Khargonekar, P., Poolla, K.: Market induced curtailment of wind power. In: 2012 IEEE Power and Energy Society General Meeting, pp. 1–8, July 2012
26. Kavanagh, R., Armstrong, D., Djemame, K., Sommacampagna, D., Blasi, L.: Towards an energy-aware cloud architecture for smart grids. In: Altmann, J., Silaghi, G.C., Rana, O.F. (eds.) GECON 2015. LNCS, vol. 9512, pp. 190–204. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43177-2_13
27. Zhang, S., Ran, Y., Wu, X., Yang, J.: Electricity cost optimization for data centers in smart grid environment. In: 11th IEEE International Conference on Control Automation (ICCA), pp. 290–295, June 2014
28. https://www.rackspace.com/managed-hosting/colocation
29. Ren, S., Islam, M.A.: Colocation demand response: why do i turn off my servers? In: 11th International Conference on Autonomic Computing (ICAC 2014), pp. 201–208. USENIX Association, Philadelphia (2014)
30. Zhan, Y., Ghamkhari, M., Xu, D., Mohsenian-Rad, H.: Propagating electricity bill onto cloud tenants: using a novel pricing mechanism. In: 2015 IEEE Global Communications Conference (GLOBECOM), pp. 1–6, December 2015

31. Zhan, Y., Ghamkhari, M., Xu, D., Ren, S., Mohsenian-Rad, H.: Extending demand response to tenants in cloud data centers via non-intrusive workload flexibility pricing. IEEE Trans. Smart Grid **99**, 1–1 (2016)
32. Christensen, H.I., Khan, A., Pokutta, S., Tetali, P.: Approximation and online algorithms for multidimensional bin packing: a survey. Comput. Sci. Rev. **24**, 63–79 (2017)