Multi-market Optimization of a Data Center without Storage Systems

Yangyang Fu[1] Wangda Zuo[1,2] Kyri Baker[1,2]

[1]Department of Civil, Architectural and Environmental Engineering, University of Colorado Boulder, USA,
{yangyang.fu,wangda.zuo,kyri.baker}@colorado.edu
[2]National Renewable Energy Laboratory, USA

Abstract

Data centers have numerous opportunities to participate in demand response programs considering their large capacities, flexible working environments and work loads, redundant design and operation, etc. Frequency regulation, as one service provided in demand response programs, can also benefits the data centers. This paper aims to develop a real-time multi-market optimization framework for a data center without storage systems to maximize their benefits from participating in both the energy market and the regulation market. Then a case study is conducted to numerically investigate the optimal bids at each hour by considering the energy cost, demand costs, and regulation revenues using a virtual data center located in PJM. Simulation results show that the proposed multi-market optimization framework can help data centers maintain minimum costs by getting maximum regulation revenues while satisfying energy and demand goals.

Keywords: Frequency Regulation, Data Center, Multi-market Optimization

1 Introduction

Data centers have numerous opportunities to participate in demand response (DR) programs considering their large energy capacities, flexible working environments and work loads, redundant design and operation, etc. For example, researches have shown that an optimized 30 MW data center is comparable to 7 MWh large-scale storage in providing DR service for the power grid (Wierman et al., 2014). Besides, some delay-tolerant data centers are allowed to have flexible work environment and workloads. What's more, the redundant design in data centers to meet reliability standards in order to guarantee their uptime and performance (Standards et al., 2005) can provide extra potentials to DR-related controls.

Frequency regulation (FR), as one type of DR, is an ancillary service that provides continuous, rapid, and automatic corrections for changes in electricity generation or use on a second-to-second basis in order to maintain the system frequency at its nominal value (e.g., 60 Hz in U.S.). Typically, FR resources are generators. FR uses certain amount of generators (e.g., about 1% of total generation) to continuously track the demand variations. The frequency must be strictly maintained within a very narrow range in order to comply with the control performance standards and the balancing authority area control error limit reliability criteria. Besides generators, fast-ramping demand side resources (DSRs) in buildings can also provide FR service to the grid by harnessing the demand flexibility provided by the modulating loads. Typical modulating loads on building side include energy storage systems such as flywheels, batteries and compressed-air energy system, electric boilers and heaters, and independent systems with variable frequency drivers (VFDs).

Recently, awareness of these potentials has drawn attention to the capabilities of data centers to participate in DR programs. A survey conducted by the Lawrence Livermore National Laboratory in 2015 shows that about 50% of the participating data centers have interest in smart pricing demand side programs, such as load shedding to avoid peak demand (Bates et al., 2015). However, data centers are reluctant to participate in fast demand response programs such as providing frequency regulation (FR) in ancillary service market, for multiple reasons. One reported concern is that data centers are still learning the process of providing FR and that providing grid services on such a fast timescale can be "outside of their visibility or control" (Bates et al., 2015). This concern is well-founded considering that these programs provide novel and relatively unexplored territory from the point of view of traditional data center control and operations.

This paper aims to explore data centers' ability of providing frequency regulation service to grids and maximize their benefits from participating regulation market and energy market as a whole. First, a synergistic control strategy together with a new regulation flexibility factor is proposed to enable the provision of regulation services in data center. Then, a real-time optimization framework is developed to maximize the data centers' benefits from participating in both the regulation market and the energy market. In Section 4, the optimization framework is evaluated in a Modelica-based environment for typical days in January and July.

2 Synergistic Strategy for Frequency Regulation

In this section, we propose a synergistic control strategy for data centers to provide FR service. This strategy is composed of four major parts. The first one is *Baseline Routine*, which predicts the baseline power usage when the data center provides no FR. The second one is *Bidding Capacity*, which is the capacity bid that the data center submits to the electrical market. The third one is *Server Power Management*, where an aggregator is adopted to represent the aggregated performance of servers in the data center. The clock frequency of the aggregator can be directly changed by a Proportional-Integral-Derivative (PID) controller in order to follow the regulation signal. Based on that, the desired frequencies for individual servers will be determined by a set of predefined assignment rules and then be propagated to all servers. The forth one is *Cooling Power Management*, which adjusts the chilled water supply temperature (CHWST) setpoint to respond to the regulation signal.

Figure 1 shows the workflow of the proposed synergistic control strategy. The *Baseline Routine* outputs the prediction of the overall power profile for the data center P_{bas} when no FR service is provided. In this paper, the prediction is performed using detailed energy models, although many other methods such as machine learning techniques can also be used. The detailed energy models and baseline settings can be referred to Section 4.1. The *Bidding Capacity* is a module that can calculate the optimal capacity bid for the data center at each time step, and output raw regulation power $\Delta P_{reg,raw}$ based on the optimal capacity bid and received regulation signal r from the electrical market. Then, the reference power P_{ref} for the data center to track is the summation of the predicted baseline power P_{bas} together with the raw regulation power $\Delta P_{reg,raw}$.

The *Server Power Management* first determines the number of required active servers in the aggregator N_{act} based on the predicted workload $\lambda^{'}$ in the next time step (e.g., one hour ahead). Then a closed-loop control using a PID controller is utilized to minimize the error between the measured total power usage P_{mea} and the reference power P_{ref} by adjusting the aggregated frequency of the server aggregator. Meanwhile, the *Cooling Power Management* applies an open-loop control to adjust the cooling system power usage by resetting the CHWST setpoint in response to the received regulation signal r.

The server aggregator receives the aggregated frequency f_{agg} and the required number of active servers N_{act} from the FR controller. Assuming there are N_0 number of servers in the data center, the server aggregator then calculates the CPU frequency f_i for an individual server i based on predefined assignment rules. The cooling system receives CHWST setpoint from the FR controller. Both the IT system and the cooling system respond in such a way that their total power P_{mea} is adjusted to track the reference power P_{ref}.

For the aggregator, there are several assignment rules to control the individual server's frequency (Li et al., 2013; Wang et al., 2019). We can also represent the aggregated server power $P_{servers}$ of all servers under an assignment rule using a simplified model (Li et al., 2013) and this approach is adopted by this paper and detailed in Section 2.1. For the FR controller, more details are described in the rest of this section.

2.1 Server Power Management

The servers in the data center can be considered as an aggregator, which is characterized by the active number of servers N_a and the aggregated frequency f. These two parameters can be determined based on the regulation signal r and incoming workload λ. The aggregrated frequency can then be distributed to the single servers as f_i using a predefined assignment algorithm. The relationship between N_a, f and r, λ is detailed in the rest of this section.

2.1.1 Server Aggregator Model

The IT equipment, especially the servers, are modelled as an aggregator, which can predict the total IT power usage and the server response time based on CPU frequency, workload arrival rate, and number of active servers (Li et al., 2013). Details are shown as follows.

$$P_{servers}(t) = \lambda(t)\sum_{0}^{r} b_i f(t)^i + \sum_{0}^{s} c_j N_a(t)^j, 0 \le i \le r, 0 \le j \le s \tag{1}$$

where b_i, c_j are constant coefficients that can be obtained from curve fitting techniques, $\lambda(t)$ is the total arrival rate, f is the aggregated relative frequency, ranging from 0 to 1, and N_a is the active number of servers at current time. f and N_a can be optimally determined in order to minimize cost.

Here we use the average response time to quantify the service quality of a data center. The workloads are modeled as GI/G/m queues, which assumes a general distribution with independent arrival times and a general distribution of service times. The total time that a job spends in the queuing system is known as response time. The response time usually consists of two parts: waiting time, that is, the time that a job spends in a queue waiting to be serviced; and service time, that is, the time that a job needs to be executed. The average response time model is adopted from (Bolch et al., 2006). Details are shown as follows.

$$\mu(t) = kf(t) \tag{2}$$

$$t_s = \frac{1}{\mu(t)} \tag{3}$$

$$\rho(t) = \frac{\lambda(t)}{N_a(t)\mu(t)}, 0 \le \rho(t) \le 1 \tag{4}$$

$$P_m = \begin{cases} \frac{\rho(t)^m + \rho(t)}{2}, & \rho(t) \ge 0.7 \\ \rho(t)^{\frac{N_a(t)+1}{2}}, & \rho(t) < 0.7 \end{cases} \tag{5}$$

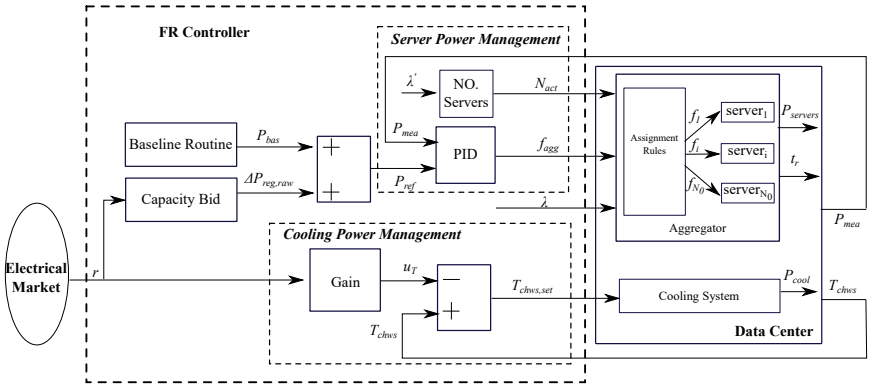

Figure 1. Data center frequency regulation control

$$t_w = \frac{C_A^2 + C_B^2}{2N_a(t)} \frac{P_m}{\mu(t)(1-\rho(t))} \tag{6}$$

$$t_r = t_s + t_w \tag{7}$$

In the above equations, μ is the mean service rate, k is a constant parameter, assuming the service rate is proportional to the frequency, ρ is the average utilization of the server, representing the fraction of occupied time, P_m is approximated probability that an arriving job is queued, C_A and C_B are constant coefficients reflecting the variations of the inter-arrival times and request sizes, t_r, t_s, and t_w are the average response time, service time and waiting time for the aggregator respectively.

2.1.2 Number of Active Servers

The number of servers in a data center needs to satisfy the following condition in order to ensure the stability of the queue. This condition means that the service rate in the data center should be greater than the arrival rate.

$$N_a(t)\mu(t) > \lambda(t) \tag{8}$$

Under design conditions, to guarantee reliability, a scaling factor γ as defined in Eq.(9) is utilized here to describe the design redundancy of the servers. The γ is set to greater than 1. If $\gamma = 1$, it means all the CPU clock frequencies need to set at maximum level just to serve the average workload, which limits the potential of FR. The γ is described as

$$\gamma = \frac{\mu_0 N_0}{\lambda_0}, \tag{9}$$

where μ_0 is the nominal service rate of a single server, N_0 is the nominal number of servers in a data center room, and λ_0 is the nominal arrival rate to be served by the data center.

When using a server aggregator model as described in Eq. (1), the γ can then be rewritten as:

$$\gamma = \frac{kN_0}{\lambda_0} = \frac{kN_a(t)}{\lambda_{mean}(t)}, \tag{10}$$

where k is a constant parameter, assuming the service rate is proportional to the aggregated frequency, N_a is the number of active servers at current time step, and λ_{mean} is the mean arrival rate at the current time step.

The number of active servers is calculated at an interval of 1 hour because the servers have relatively long wakeup time. The detailed formula is shown in Eq. (11), where the operator $\lceil x \rceil$ is the ceiling function which yields the smallest integer greater or equal to x.

$$N_a(t) = \lceil \frac{\gamma \lambda_{mean}(t)}{k} \rceil \tag{11}$$

By adding a FR flexibility factor β during operation, we can determine the number of active servers based on the predicted coming arrival rate, as shown in Eq. (12). The greater β is, the more servers are activated for a specific workload.

$$N_a(t) = \lceil \beta \frac{\gamma \lambda_{mean}(t)}{k} \rceil, N_a(t) \in [0, N_0] \tag{12}$$

2.1.3 Frequency Control

The aggregated frequency f_{agg} is controlled by a PID controller to track the reference power P_{ref} calculated from the electrical market. The reference power P_{reg} is calculated as

$$\Delta P_{reg,raw}(t) = r(t) C_{reg} \tag{13}$$

$$P_{ref}(t) = P_{bas}(t) + \Delta P_{reg,raw}(t) \tag{14}$$

where $\Delta P_{reg,raw}$ is the raw power signal and C_{reg} is the regulation capacity that the data center bids in the market.

The frequency f_{agg} is then determined by the PID controller as follows.

$$f_{agg}(t) = K_p e(t) + K_i \int_0^t e(x)dx + K_d \frac{\mathrm{d}e(t)}{\mathrm{d}t}, f_{agg}(t) \in [f_{min}, f_{max}] \tag{15}$$

$$e(t) = P_{ref}(t) - P_{mea}(t) \tag{16}$$

In the above equations, K_p, K_i, and K_d denote the coefficients for the term P, I and D, respectively. e is the error between the reference power P_{ref} and the measured power

P_{mea}. The maximum aggregated frequency is 1, while the minimum frequency varies based on the number of active servers due to the constraints of Quality of Service (QoS). Details on how to determine f_{min} are described in Section 2.1.4.

2.1.4 Minimum Aggregate Frequency

Using a service response time model shown in Eq. (2) and Eq. (7), we know that the response time of the servers depends on the aggregated frequency. If the frequency is low, then it takes relatively long time for the servers to respond to the arrival workload, which means the QoS of the data center is compromised. To enable FR and guarantee the QoS, the aggregated frequency should meet a minimum value. The minimum can be obtained by solving the following optimization problem.

$$\begin{aligned} \min \quad & f(\rho(t)) = \frac{\lambda(t)}{kN_a(t)\rho(t)} \\ s.t. \quad & 0 \le \rho(t) \le 1 \\ & t_r(t) \le t_{r,u} \end{aligned} \tag{17}$$

where $\rho(t)$ is the utilization rate as defined in Eq. (4), $t_r(t)$ is the service response time as calculated in Eq. (7) and t_u is the maximum response time allowed by the data center.

Rearranging Eq. (2) to Eq. (7), we can get the response time $t_r(t)$ as a function of the utilization rate $\rho(t)$ as follows.

$$t_r(\rho(t)) = \frac{\rho(t)}{\lambda(t)}[N_a(t) + \frac{C_A^2 + C_B^2}{2(1-\rho(t))}P_m(\rho(t))] \tag{18}$$

It is easy to show that

$$\frac{\mathrm{d}t_r(\rho)}{\mathrm{d}\rho} > 0 \tag{19}$$

Thus, the above-mentioned optimization problem can be solved at each time step as:

$$f_{min}(t) = \frac{\lambda(t)}{kN_a(t)\rho^*(t)} \tag{20}$$

where $\rho^*(t)$ is the optimal utilization rate, and $\rho^*(t)$ should satisfy the nonlinear relationship shown as:

$$t_r(\rho^*(t)) - t_u = 0 \tag{21}$$

2.2 Cooling Power Management

The cooling system power is managed by resetting the chilled water supply temperature. The regulation signal from the electrical market is directly used to change the chilled water supply temperature setpoint $T_{chws,set}$ by Eq. (22).

$$T_{chws,set}(t) = T_{chws}(t) - \Delta T r(t) \tag{22}$$

where T_{chws} is the chilled water temperature at current time step, ΔT is the user defined regulation range for the temperature, and varies based on the design supply temperature range of chillers. Here we set it to 2 °C. The negative sign at the right term means when regulation up is needed, the temperature setpoint is reduced, and vice versa.

3 Multi-market Optimization Framework

A real-time optimization framework is applied for optimizing the operation of the data center without thermal storage system in the presence of real-time (or day-ahead) energy prices, peak demand charges, and frequency regulation revenue. For each optimization time step, the overall objective can be described as:

$$\begin{aligned} \min \quad & J(C_{reg}) = E_{cost} + D_{cost} - R_{revenue} \\ s.t. \quad & 0 \le C_{reg}(t) \le C_{reg,max}(t) \\ & t_r(t) \le t_{r,u} \\ & S(C_{reg}) \ge S_l \end{aligned} \tag{23}$$

where C_{reg} is the design variable, representing regulation capacity bid at each hour, $C_{reg,max}$ is the maximum capacity the data center can provide for regulation, t_r is the response time of the data center service, $t_{r,u}$ is the allowable upper limit of the response time, S is the regulation performance score defined by PJM as shown in Section 6.1, and S_l is the lowest allowable performance score by PJM to participate in regulation market.

The cost function J has three terms: energy cost E_{cost}, demand cost D_{cost} and regulation revenue $R_{revenue}$. The energy cost is calculated by Eq. (24).

$$E_{cost} = \int_t^{t+\Delta t} p_{em}(t) P_{DC}(t) dt \tag{24}$$

where p_{em} is the real-time price signals for energy use at time t, P_{DC} is the total power consumption for the data center at time t. The calculation period starts from time t and ends at $t + \Delta t$, where Δt is the optimization step, and is set to 1 hour in this study.

The electric demand during the current optimization horizon is penalized by the demand price p_{dm} as shown in Eq. (25).

$$D_{cost} = p_{dm} \cdot \max((P_{dm} - P_{dm,lim}), 0) \tag{25}$$

where p_{dm} is the demand price, P_{dm} is the power demand calculated as the average power for each 30-min interval, and $P_{dm,lim}$ is the limit of required demand. This function means if the demand in current step exceeds a predefined demand value, then the optimization cost function is penalized by the demand difference. Otherwise, no penalization is applied. Note that p_{dm} and P_{dm} are both utility specific, and may vary from this definition.

The revenues from regulation service is computed as follows.

$$R_{revenue} = \int_{t}^{t+\Delta t} p_{rm}(t) C_{reg}(t) dt \tag{26}$$

where p_{rm} is the real-time price signal from the regulation market, and $C_{reg}(t)$ is the regulation capacity bid for each time step.

The price signals such as p_{em} and p_{rm} need to be predicted one optimization step ahead, e.g. 1 hour in this study. Many researches have been conducted for this purpose. In this paper, historical prices of these two electrical markets are used, which means the hourly ahead prices are assumed to be perfectly predicted. The demand limit $P_{dm,lim}$ can also be predefined by the data center operators based on historical operation conditions. The maximum regulation capacity at each optimization step is set to 798.2 kW (20% of the nominal power). Note this maximum regulation capacity setting is not the feasible capacity the data center can provide at each hour, because the regulation capacity is related to data center operational conditions such as arrival rate, and weather conditions etc. This simplification has limited influence on the optimization results when the lower limit of performance score s_l is set to a high value, because if the data center makes a bid that exceeds its capacity, it cannot track the reference signal, thus the regulation performance will be low. By setting s_l to a high value can help data center make a reasonable bids when the regulation capacity is hard to predict. The optimization problem is solved using the pattern search algorithm in the optimization engine, GenOpt (Wetter et al., 2001).

4 Case Study

A data center as shown in Figure 2 is used to investigate the benefits from participating in different electrical markets. The data center is considered as a price taker only. This case study investigates the maximum benefits that data centers can obtain from both the real-time energy market and the regulation market in PJM. For the regulation service, only dynamic regulation is studied here, because its price is usually much higher than traditional regulation.

4.1 Case Description

The data center is located in Chicago, which is in ASHRAE Climate Zone 5A and within the PJM market territory. For the cooling system, there are two chillers and one integrated waterside economizer providing cooling to the data center room. This cooling system can operate in three modes: Free Cooling (FC) mode when only the WSE is enabled for cooling, Partial Mechanical Cooling (PMC) mode when the chiller and WSE are both triggered, and Full Mechanical Cooling (FMC) mode when only the chiller is activated. There are also two cooling towers, two constant-speed condenser water pumps, two variable-speed chilled water pumps, and one variable speed fan. The cooling system and its control are modelled using

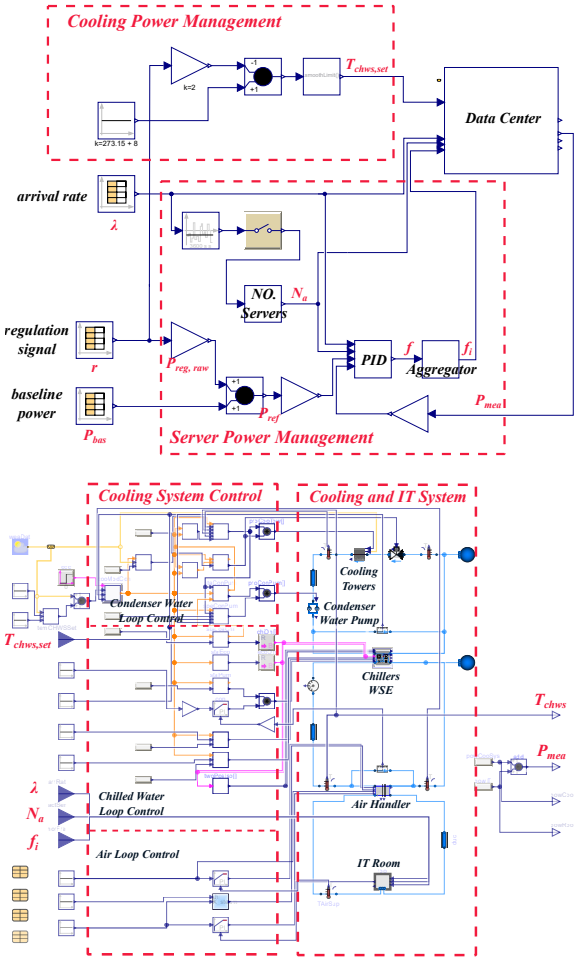

Figure 2. Modelica implementation of the studied data center for FR service: FR controller (top) and data center system (bottom)

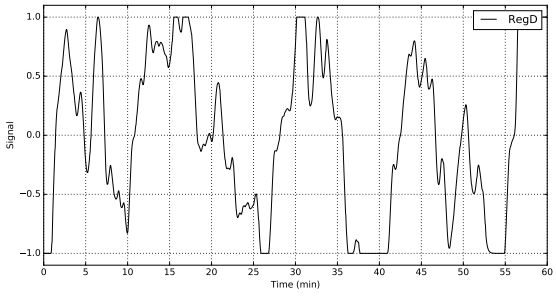

Figure 3. An example of one-hour historical RegD signal in January

an open-source equation-based Modelica environment (Fu et al., 2018, 2019a,c,b).

For the IT system, the design number of servers is 8000. The design factor γ is set to 1.5 (Li et al., 2013). The total nominal electrical load is about 2700 kW. The calibrated coefficients for Eq. (1) are $b_0 = 0.0154$, $b_1 = 1.5837$, $b_2 = 0.1373$, $c_0 = -22.3540$ and $c_1 = 121.0212$ using the method mentioned in Ref. (Li et al., 2013). When not providing FR, the server aggregator operates at a frequency of 0.8 with a regulation flexibility factor of 1.0, and the CHWST setpoint is set to 8 °C. For the internet data center, the constants C_A and C_B are set to 1 as in Ref. (Li et al., 2013).

For the multi-market optimization, all the settings are the same as the baseline except that an additional FR controller as designed in Section 2 is used to provide regulation service for the grids by adjusting the CPU frequency and CHWST setpoint. The FR flexibility factor is set to 1.1 when providing regulation services. The QoS when providing regulation services is guaranteed by constraining the average response time of the data center service to 6 ms. The lower limit of the performance score in PJM to disqualify a regulation resource is 0.4 (LLC, 2019). Here we set it to a higher value, 0.9. The real-time optimization is performed at a one-hour interval for 2 days in both January ($1/20 \sim 1/21$) (when cooling system operates at FC mode) and July ($7/20 \sim 7/21$) (when cooling system operates at FMC mode).

The price signals of the real-time energy market and the regulation service market in January and July 2018 are posted in Ref. (PJM, 2019), and the price during the optimization period is plotted as shown in Figure 4. An example of one-hour historical RegD signal is plotted in Figure 3. A real-time web service in Wikipedia (Wang et al., 2019) is used as the workload arrival profile during optimization, which is shown in Figure 5.

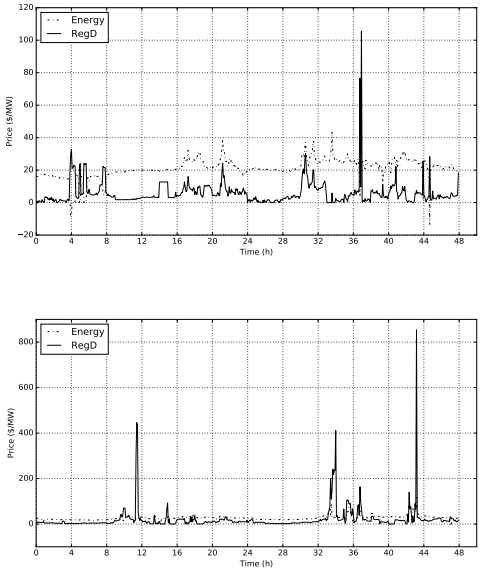

Figure 4. Historical real-time prices of PJM energy market and regulation market in January (top) and July (bottom)

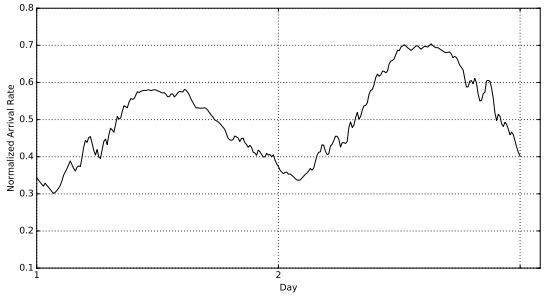

Figure 5. Two-day historical arrival rates in the data center

4.2 Results and Discussions

Table 1 compares the total cost of the data center in terms of baseline operation and multi-market optimization. The baseline system is denoted as *Base*, and the multi-market optimization is denoted as *OPT*. In both January and July, the data center without energy storage systems, using the proposed optimization framework, can benefit from participating in both energy market and regulation market. In the two days considered, *OPT* can save $123.6 in July, while the saving is $24.8 in January.

The savings mainly come from the revenues in the regulation market, and the cost for energy use and demand charge are almost the same in the *Base* and *OPT*. Because the sum of the RegD signal over a long time period (e.g. 1 hour) is almost 0, providing regulation service in the *OPT* leads to the similar energy use, thus similar energy cost compared with the *Base* where no regulation service is provided. By utilizing the demand cost defined in Eq. (25), the data center can provide regulation service without increasing monthly demand, thus no extra demand

Table 1. Multi-market optimization of data centers

Costs	January		July	
	Base	*OPT*	*Base*	*OPT*
Energy Cost ($)	1043.3	1042.9	1591.4	1590.6
Demand Cost ($)	10459.3	10457.5	12063.9	12062.8
Regulation Revenue ($)		22.6		121.7
Total Cost ($)	11502.6	11477.8	13655.3	13531.7
Total Savings ($)		24.8		123.6

charge would be added to utility bills. The revenue from July is much higher than that in January because the price for dynamic regulation (RegD) resources is higher in July. As shown in Figure 4, during the studied two days, the average price from regulation market in July is about 21 $/MW, while that in January is only about 5.8 $/MW.

Figure 6 shows the hourly capacity bids in 1/21 and 7/21. The demand for each 30 minutes is denoted as the thin solid line. The demand limit used for demand cost as shown in Eq. (25) is denoted as the dashed line. The optimal capacity bid at each hour is denoted as the shaded area. At non-peak hours (e.g., 3:00 - 6:00), the optimal bid is mainly influenced by the price from energy market, price from regulation market and detailed shape of RegD signal. Because the demand is lower than the demand limit, the tradeoff between the energy cost and revenues from regulation market determines the optimal bid. The energy cost is highly influenced by the energy use, which is determined by the detailed shape of the RegD signal. If the sum of the RegD signal is larger than 0, then more energy would be consumed when providing frequency regulation service, thus the energy cost would increase. Although the energy cost increases in this case, the data center can get revenues from regulation market. If the sum of the RegD signal is no larger than 0, then at that hour, the data center can bid at their maximum capacity.

At peak hours (e.g., 12:00 - 16:00), the optimal bid is mostly influenced by the demand limit and the RegD signal. Figure 6 shows that at these hours, the bid is small so that the demand cannot exceed the required demand limit to avoid demand penalty. At 13:00, the bid is about 69 kW, but it is only about 5 kW at 14:00. The difference is caused by the detailed shapes of the RegD signals in these two hours. At 13:00, the sum of the RegD signal in first 30 minutes is slightly greater than 0, but in the second 30 minutes it is much smaller than 0. This means that regulation capacity bid in this hour can increase the demand in the first 30 minutes, but the demand in the second 30 minutes can be decreased compared with the same time in the baseline system. Therefore, at this hour, the data center can bid a large capacity as long as the demand in the first 30 minutes will not exceed the demand limit. The same situation happens at 14:00 but with a large sum of RegD signal at first 30 minutes. Also because the power at 14:00 is much closer to the demand limit, the data center can only bid a small capacity at this hour.

In summary, the proposed real-time optimization framework can help the data center without energy storage system harness the benefits from the energy market and the regulation market. However, the benefits are insignificant compared with the large baseline power in data centers. One of the reason is that data centers without energy storage system are difficult to limit their power demand during FR service, which contributes to a large portion of the utility bill. In the future, we will consider retrofit strategy (e.g., installing thermal storage energy system) in the data center to limit the power demand to maximize the benefit from the multi-markets

5 Conclusions

This paper developed a real-time multi-market optimization framework for the data center without storage systems to maximize their benefits from participating in both energy market and regulation market. Then, a case study was conducted to numerically investigate the optimal bids at each hour by considering the energy cost, demand costs and regulation revenues using a virtual data center located in PJM. Simulation results shows that using the proposed multi-market optimization framework can minimize the operational cost. Compared with the baseline system, providing frequency regulation service over the considered two days can save $24.8 in January and $123.6 in July.

6 Appendix

6.1 FR Performance Score

In the PJM market, new resources aiming to enter the regulation market need to pass an initial test by obtaining at least 0.75 for a defined performance score. The initial test signals of RegA and RegD are available at (PJM, 2019). The performance score is calculated as a composite score of accuracy, delay and precision, which are shown below (LLC, 2019).

$$c_{sig,res} = \frac{COV(reg, res)}{\sigma_{reg}\sigma_{res}} \tag{27}$$

$$S_{accuracy} = \max_{\delta=0-5\ \text{min}} (c_{reg,res(\delta)}) \tag{28}$$

$$S_{delay} = \left| \frac{5\ \text{min} - \delta^*}{5\ \text{min}} \right| \tag{29}$$

$$S_{precision} = 1 - \frac{1}{n} \sum \left| \frac{res - reg}{\overline{reg}} \right| \tag{30}$$

$$S = \frac{S_{accuracy} + S_{delay} + S_{precision}}{3} \tag{31}$$

In the above equations, *reg* represents the regulation signal the DSRs receive from the electrical markets, and *res* represents the response signal the DSRs generate after control actions. c, COV and σ are the correlation coefficient, covariance, standard deviation of these two signals. In PJM, the response signal *res* is recalculated with a time shift δ ranging from 0 to 5 minutes in an increment of 10 seconds, which leads to 31 response signals $res(\delta)$. The accuracy score $S_{accuracy}$ is the maximum correlation coefficient c between *reg* and $res(\delta)$. The delay score S_{delay} is calculated based on the delay time δ^* when the maximum accuracy score is obtained using Eq. (29). The precision score $S_{precision}$ is defined as the relative difference between regulation signal and response signal, where n is the number of samples in the hour, and $\overline{reg}$ is the hourly average regulation signal. The final performance score S in that hour is calculated as the weighted average of the three individual scores.

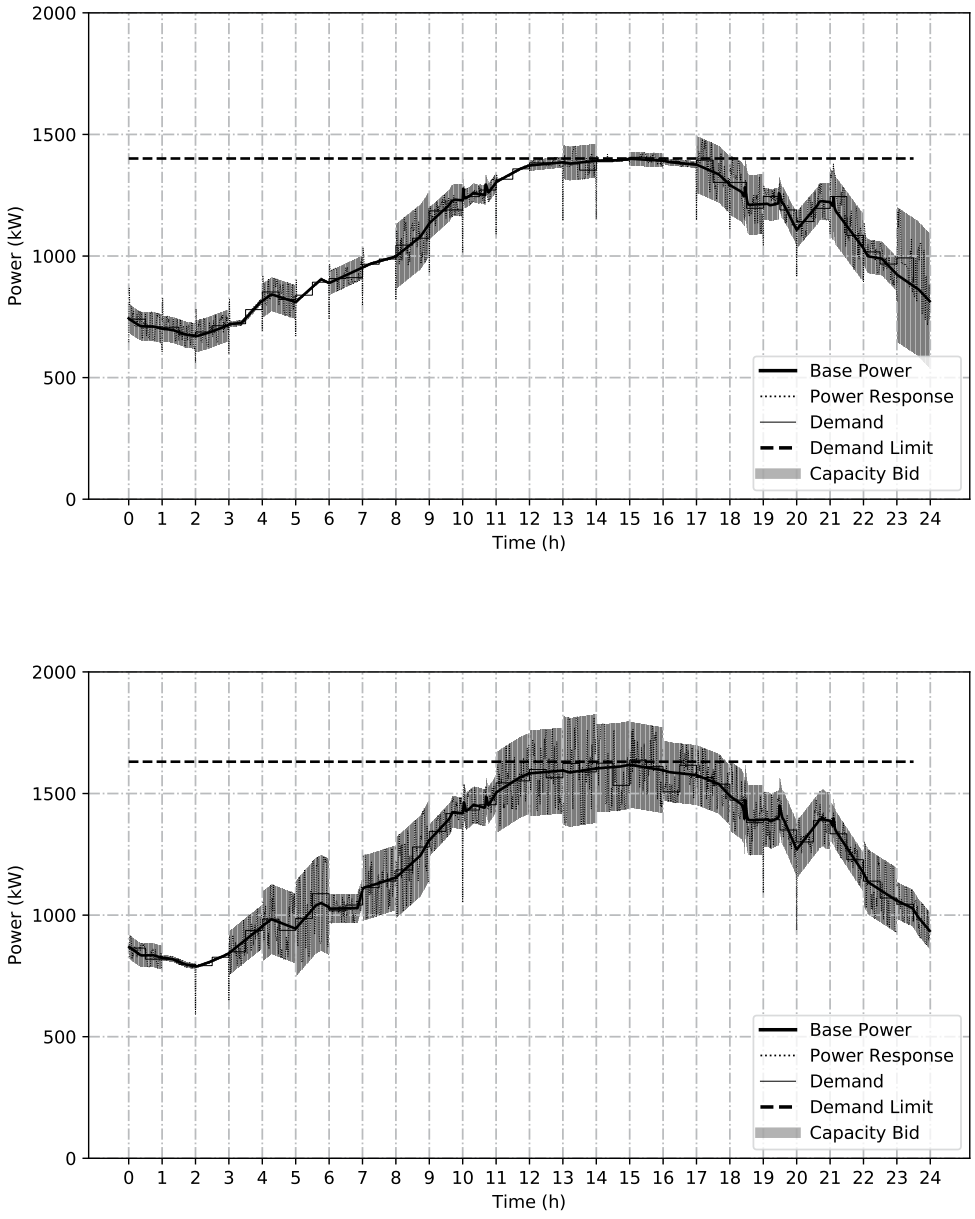

Figure 6. Optimal hourly regulation capacity bids in 1/21 (top) and 7/21 (bottom)

References

Natalie Bates, Girish Ghatikar, Ghaleb Abdulla, Gregory A Koenig, Sridutt Bhalachandra, Mehdi Sheikhalishahi, Tapasya Patki, Barry Rountree, and Stephen Poole. Electrical grid and supercomputing centers: An investigative analysis of emerging opportunities and challenges. *Informatik-Spektrum*, 38(2):111–127, 2015.

Gunter Bolch, Stefan Greiner, Hermann De Meer, and Kishor S Trivedi. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons, 2006.

Yangyang Fu, Michael Wetter, and Wangda Zuo. Modelica models for data center cooling systems. In *2018 Building Performance Analysis Conference and SimBuild, Chicago, Illinois, United States of America*, 2018.

Yangyang Fu, Xing Lu, and Wangda Zuo. Modelica models for the control evaluations of chilled water system with waterside economizer. In *Proceedings of the 13th International Modelica Conference, Regensburg, Germany, March 4ñ6, 2019*, number 157, page 8. Link^ping University Electronic Press, Link^pings universitet, 2019a.

Yangyang Fu, Wangda Zuo, Michael Wetter, James W. VanGilder, and Peilin Yang. Equation-based object-oriented modeling and simulation of data center cooling systems. *Energy and Buildings*, 198:503 – 519, 2019b. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2019.06.037. URL `http://www.sciencedirect.com/science/article/pii/S0378778819307078`.

Yangyang Fu, Wangda Zuo, Michael Wetter, Jim W. VanGilder, Xu Han, and David Plamondon. Equation-based object-oriented modeling and simulation for data center cooling: A case study. *Energy and Buildings*, 186:108 – 125, 2019c. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2019.01.018. URL `http://www.sciencedirect.com/science/article/pii/S0378778818330573`.

Sen Li, Marco Brocanelli, Wei Zhang, and Xiaorui Wang. Data center power control for frequency regulation. In *2013 IEEE Power & Energy Society General Meeting*, pages 1–5. IEEE, 2013.

PJM Interconnection LLC. Pjm manual 11: Energy & ancillary services market operations, 2019.

PJM. Ancillary services, 2019. URL `https://www.pjm.com/markets-and-operations/ancillary-services.aspx`.

Telecommunication Industry Association. Standards, Technology Dept, and American National Standards Institute. *Telecommunications Infrastructure Standard for Data Centers*. Telecommunication Industry Association, 2005.

Wei Wang, Amirali Abdolrashidi, Nanpeng Yu, and Daniel Wong. Frequency regulation service provision in data center with computational flexibility. *Applied Energy*, 251:113304, 2019.

Michael Wetter et al. Genopt-a generic optimization program. In *Seventh International IBPSA Conference, Rio de Janeiro*, pages 601–608, 2001.

Adam Wierman, Zhenhua Liu, Iris Liu, and Hamed Mohsenian-Rad. Opportunities and challenges for data center demand response. In *International Green Computing Conference*, pages 1–10. IEEE, 2014.