Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
(2013) 1–28

Data Center Demand Response: Avoiding the Coincident Peak via Workload Shifting and Local Generation

Zhenhua Liu[1], Adam Wierman[1], Yuan Chen[2], Benjamin Razon[1], Niangjun Chen[1]

[1]*California Institute of Technology* [2]*HP Labs*

[1]{*zhenhua,adamw,ben,ncchen*}*@caltech.edu* [2]*yuan.chen@hp.com*

Abstract

Demand response is a crucial aspect of the future smart grid. It has the potential to provide significant peak demand reduction and to ease the incorporation of renewable energy into the grid. Data centers' participation in demand response is becoming increasingly important given their high and increasing energy consumption and their flexibility in demand management compared to conventional industrial facilities. In this paper, we study two demand response schemes to reduce a data center's peak loads and energy expenditure: workload shifting and the use of local power generation. We conduct a detailed characterization study of coincident peak data over two decades from Fort Collins Utilities, Colorado and then develop two algorithms for data centers by combining workload scheduling and local power generation to avoid the coincident peak and reduce the energy expenditure. The first algorithm optimizes the expected cost and the second one provides a good worst-case guarantee for any coincident peak pattern, workload demand and renewable generation prediction error distributions. We evaluate these algorithms via numerical simulations based on real world traces from production systems. The results show that using workload shifting in combination with local generation can provide significant cost savings (up to 40% under the Fort Collins Utilities charging scheme) compared to either alone.

1. Introduction

Demand response (DR) programs seek to provide incentives to induce dynamic demand management of customers' electricity load in response to power supply conditions, for example, reducing their power consumption in response to a peak load warning signal or request from the utility. The National Institute of Standards and Technology (NIST) and the Department of Energy (DoE) have both identified demand response as one of the priority areas for the future smart grid [1, 2]. In particular, the National Assessment of Demand Response Potential report has identified that demand response has the potential to reduce up to 20% of the total peak electricity demand across the country [3]. Further, demand response has the potential to significantly ease the adoption of renewable energy into the grid.

Data centers represent a particularly promising industry for the adoption of demand response programs. First, data center energy consumption is large and increasing rapidly. In 2011, data centers consumed approximately 1.5% of all electricity worldwide, which was about 56% higher than the preceding five years [4, 5, 6, 7]. Second, data centers are highly automated and monitored, and so there is the potential for a high-degree of responsiveness. For example,

today's data centers are well instrumented with a rich set of sensors and actuators. The power load and state of IT equipment (e.g., server, storage and networking devices) and cooling facility (e.g., CRAC units) can be continuously monitored and panoramically adjusted. Third, many workloads in data centers are delay tolerant, and can be scheduled to finish anytime before their deadlines. This enables significant flexibility for managing power demand. Finally, local power generation, e.g., both traditional backup generators such as diesel or natural gas powered generators and newer renewable power installations such as solar PV arrays, can help reduce the need from the grid by supplying the demand at critical times. In particular, local power generation combined with workload management has a significant potential to shed the peak load and reduce energy costs.

Despite wide recognition of the potential for demand response in data centers, the current reality is that industry data centers seemingly perform little, if any, demand response [4, 5]. One of the most common demand response programs available is Coincident Peak Pricing (CPP), which is required for medium and large industrial consumers, including data centers, in many regions. These programs work by charging a very high price for usage during the coincident peak hour, often over 200 times higher than the base rate, where the coincident peak hour is the hour when the most electricity is requested by the utility from its wholesale electric supplier. It is common for the coincident peak charges to account for 23% or more of a customer's electric bill according to Fort Collins Utilities [8]. Hence, from the perspective of a consumer, it is critical to control and reduce usage during the peak hour. Although it is impossible to accurately predict exactly when the peak hour will occur, many utilities identify potential peak hours and send warning signals to customers, which helps customers manage their loads and make decisions about their energy usage. For example, Fort Collins Utilities sends coincident peak warnings for 3-22 hours each month with average 14.5 in summer months and 10 in winter ones. Depending on the utility, warnings may come between 5 minutes and 24 hours ahead of time.

Coincident peak pricing is not a new phenomenon. In fact, it has been used for large industrial consumers for decades. However, it is rare for large industrial consumers to have the responsiveness that data centers can provide. Unfortunately, data centers today either do not respond to coincident peak warnings or simply respond by turning on their backup power generators [9]. Using backup power generation seems appealing since it can be automated easily, it does not impact operations, and it provides demand response for the utility company. However, the traditional backup generators at data centers can be very "dirty" – in some cases even not meeting Environmental Protection Agency (EPA) emissions standards [4]. So, from an environmental perspective this form of response is far from ideal. Further, running a backup generator can be expensive. Alternatively, providing demand response via shifting workload can be more cost effective. One of the challenges with workload shifting is that we need to ensure that the Service Level Agreements (SLAs), e.g., completion deadlines, remain satisfied even with uncertainties in coincident peak and warning patterns, workload demand, and renewable generation.

1.1. Summary of contributions

Our main contributions are the following. First, we present **a detailed characterization study of coincident peak pricing** and provide insight about its properties. Section 2 discusses the characterization of 26 years' coincident peak pricing data from Fort Collins Utilities in Colorado. The data highlights a number of important observations about coincident peak pricing (CPP). For example, the data set shows that both the coincident peak occurrence and warning occurrence have strong diurnal patterns that differ considerably during different days of the week and seasons. Further,

the data highlights that coincident peak warnings are highly reliable – only twice did the coincident peak not occur during a warning hour. Finally, the data on coincident peak warnings highlights that the frequency of warnings tends to decrease through the month, and that there tend to be less than seven days per month on which warnings occur.

Second, we develop **two algorithms for avoiding the coincident peak and reducing the energy expenditure using workload shifting and local power generation.** Though there has been considerable recent work studying workload planning in data centers, e.g., [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20], the uncertainty of the occurrence of the coincident peak hour presents significant new algorithmic challenges beyond what has been addressed previously. In particular, small errors in the prediction of workload or renewable generation have only a small effect on the resulting costs of workload planning; however, errors in the prediction of the coincident peak have a threshold effect – if you are wrong you pay a large additional cost. This lack of continuity is well known to make the development of online algorithms considerably more challenging.

Given the challenges associated with the combination of uncertainty about the coincident peak hour and warning hours, workload demand, and renewable generation, we consider two design goals when developing algorithms: good performance in the average case and in the worst case. We develop an algorithm for each goal. For the average case, we present a stochastic optimization based algorithm given the estimates of the likelihood of a coincident peak or warning during each hour of the day, and predictions of workload demand and renewable generation. The algorithm provides provable robustness guarantees in terms of the variance of the prediction errors. For the worst case scenario, we propose a robust optimization based algorithm that is computationally feasible and simple, and guarantees that the cost is within a small constant of the optimal cost of an offline algorithm for any coincident peak and warning patterns, workload demand, and renewable generation prediction error distributions with bounded variance. Note that a distinguishing feature of our analysis is that we provide provable bounds on the impact of prediction errors. In prior work on data center capacity provisioning prediction errors have almost always been studied via simulation, if at all.

The third main contribution of our work is **a detailed study and comparison of the potential cost savings of algorithms via numerical simulations based on real world traces from production systems**. The experimental results in Section 5 highlight a number of important observations. Most importantly, the results highlight that our proposed algorithms provide significant cost and emission reductions compared to industry practice and provide close to the minimal costs under real workloads. Further, our experimental results highlight that both local generation and workload shifting are important for ensuring minimal energy costs and emissions. Specifically, combining workload shifting with local generation can provide 35-40% reductions of energy costs, and 10-15% reductions of emissions. We also illustrate that our algorithms are robust to prediction errors.

1.2. Related work

While the design of workload planning algorithms for data centers has received considerable attention in recent years, e.g., [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20] and the references therein; demand response for data centers is a relatively new topic. Some of the initial work in the area comes from Urgaonkar et al. [21], which proposes an approach for providing demand response by using energy storage to shift peak demand away from high peak periods. This technique complements other demand response schemes such as workload shifting. Conceptually, using local storage is similar to the use of local power generation studied in the current paper. In this paper, we consider both the workload shifting and local power generation. The integration of energy storage to our framework is a topic of our

future work. Another recent approach for data center demand response is Irwin et al. [22], which studies a distributed storage solution for demand response where compatible storage systems are used to optimize I/O throughput, data availability, and energy-efficiency as power varies. Perhaps the most in depth study of data center demand response to this point is the recent report released by Lawerence Berkeley National Laboratories (LBNL) [5]. This report summarizes a field study of four data centers and evaluates the potential of different approaches for providing demand response. Such approaches include adjusting the temperature set point, shutting down or idling IT equipment and storage, load migration, and adjusting building properties such as lighting and ventilation. The results show that data centers can provide 10-12% energy usage savings at the building level with minimal or no impact to data center operations. This report highlights the potential of demand response and shows that it is feasible for a data center to respond to signals from utilities, but stops short of providing algorithms to optimize cost in demand response programs, which is the focus of the current paper.

2. Coincident peak pricing

Most typically, the demand response programs available for data centers today are some form of coincident peak pricing. In this section, we give an overview of coincident peak pricing programs and then do a detailed characterization of the coincident peak pricing program run by Fort Collins Utilities in Colorado, where HP has a data center charged by this company.

2.1. An overview of coincident peak pricing

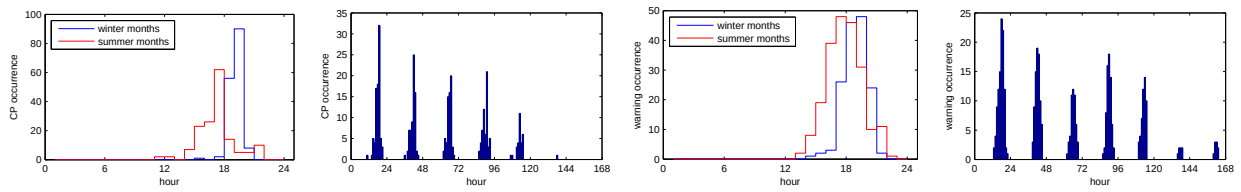
In a coincident peak pricing program, a customer's monthly electricity bill is made up of four components: (i) a fixed connection/meter charge, (ii) a usage charge, (iii) a peak demand charge for usage during the customer's peak hour, and (iv) a coincident peak demand charge for usage during the coincident peak (CP) hour, which is the hour during which the utility company's usage is the highest. Each of these is described in detail below.

Connection/Meter charge. The connection and meter charges are fixed charges that cover the maintenance and construction of electric lines as well as services like meter reading and billing. For medium and large industrial consumers such as data centers, these charges make up a very small fraction of the total power costs.

Usage charge. The usage charge in CPP programs works similarly to the way it does for residential consumers. The utility specifies the electricity price $\$p(t)$/kWh for each hour. This price is typically fixed throughout each season, but can also be time-varying. Usually $p(t)$ is on the order of several cents per kWh.

Peak demand charge. CPP programs also include a peak demand charge in order to incentivize customers to consume power in a uniform manner, which reduces costs for the utility due to smaller capacity provisioning. The peak demand charge is typically computed by determining the hour of the month during which the customer's electricity use is highest. This usage is then charged at a rate of $\$p_p$/kWh, which is much higher than $p(t)$. It is typically on the order of several dollars per kWh.

Coincident peak charge. The defining feature of CPP programs is the coincident peak charge. This charge is similar to the peak charge, but focuses on the peak hour for the utility as a whole from its wholesale electricity provider (the coincident peak) rather than the peak hour for an individual consumer. In particular, at the end of each month the peak usage hour for the utility, t_{cp}, is determined and then all consumers are charged $\$p_{cp}$/kWh for their usage



(a) Time of the day (b) Days of the week (c) Time of the day (d) Days of the week

Figure 1. Occurrence of coincident peak and warnings. (a) Empirical frequency of CP occurrences on the time of day, (b) Empirical frequency of CP occurrences over the week, (c) Empirical frequency of warning occurrences on the time of day, and (d) Empirical frequency of warning occurrences over the week.

during this hour. This rate is again at the scale of several dollars per kWh, and can be significantly larger than the peak demand charging rate p_p.

Note that customers cannot know when the coincident peak will occur since it depends on the behavior of all of the utility's customers. As a result, to aid customers the utility sends *warnings* that particular hours may be *the* coincident peak hour. Depending on the utility, these warnings can be anywhere from 5 minutes to 24 hours ahead of time, though they are most often in the 5-10 minute time-frame. These warnings can last multiple hours and can occur anywhere from two to tens of times during a month. In practice, these warnings are extremely reliable – the coincident peak almost never occurs outside of a warning hour. This is important since warnings are the only signal the utility has for achieving responsiveness from customers.

2.2. A case study: Fort Collins Utilities Coincident Peak Pricing (CPP) Program

In order to provide a more detailed understanding of CPP programs, we have obtained data from the Fort Collins Utilities on the CPP program they run for medium and large industrial and commercial customers. The data we have obtained covers the operation of the program from January 1986 to June 2012. It includes the date and hour of the coincident peak each month as well as the date, hour, and length of each warning period. In the following we focus our study on three aspects: the rates, the occurrence of the coincident peak, and the occurrence of the warnings.

Rates. We begin by summarizing the prices for each component of the CPP program. The rates for 2011 and 2012 are summarized in the Table 1. It is worth making a few observations. First, note that all the prices are fixed and announced at the beginning of the year, which eliminates any uncertainty about prices with respect to data center planning. Further, the prices are constant within each season; however the utility began to differentiate between summer months and winter months in 2012. Second, because the coincident peak price and the peak price are both so much higher than the usage price, the costs associated with the coincident peak and the peak are important components of the energy costs of a data center. In particular, $\frac{p_p}{p}$ is 194 and 148, and $\frac{p_{cp}}{p}$ is 514 and 219, in 2011 and winter 2012, respectively. Hence, it is very critical to reduce both the data center peak demand and the coincident peak demand in order to lower the total cost. A final observation is that the coincident peak price is higher than the peak demand price, 2.6 times and 1.4 times higher in 2011 and winter 2012, respectively. This means that the reduction of power demand during the coincident peak hour is more important, further highlighting the importance of avoiding coincident peaks.

Coincident peak. Understanding properties of the coincident peaks is particularly important when considering data center demand response. Figure 1 summarizes the coincident peak data we have obtained from Fort Collins Utilities from January 1986 to June 2012. Figure 1(a) depicts the number of coincident peak occurrences during each

Charging rates	2011	2012
Fixed $/month	54.11	61.96
Additional meter $/month	47.81	54.74
CP summer $/kWh	12.61	10.20
CP winter $/kWh	12.61	7.64
Peak $/kWh	4.75	5.44
Energy summer $/kWh	0.0245	0.0367
Energy winter $/kWh	0.0245	0.0349

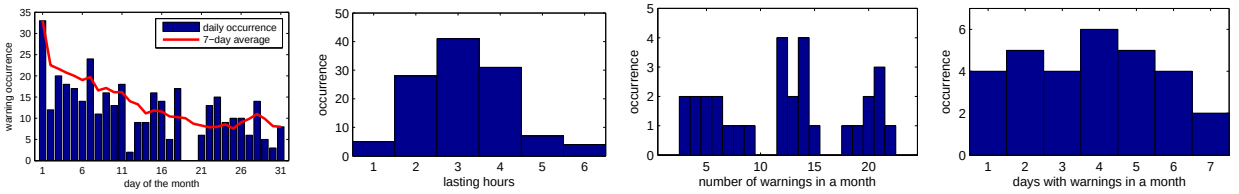
Table 1. Summary of the charging rates of Fort Collins Utilities during 2011 and 2012 [8].

hour of the day. From the figure, we can see that the coincident peak has a strong diurnal pattern: the coincident peak nearly always happens between 2pm and 10pm. Additionally, the figure highlights that the coincident peak has different seasonal patterns in winter and summer: the coincident peak occurs later in the day during winter months than during summer months. Further, the time range that most coincident peaks occur is narrower during winter months. The number of coincident peak occurrences on a weekly basis is shown in Figure 1(b). The data shows that the coincident peak has a strong weekly pattern: the coincident peak almost never happens on the weekend, and the likelihood of occurrence decreases throughout the weekdays.

Warnings. To facilitate customers managing their demand, Fort Collins Utilities identify potential peak hours and send warning signals to customers. These warnings are the key tool through which utilities achieve responsiveness from customers, i.e., demand response. On average, warnings from Fort Collins Utilities cover 12 hours for each month. Figures 1(c), 1(d), and 2 summarize the data on warnings announced by Fort Collins Utilities between January 2010 and June 2012. We limit our discussion to this period because the algorithm for announcing warnings was consistent during this interval. During this period, warnings were announced 5-10 minutes before the warning period began. Note that warnings are only useful if they do in fact align with the coincident peak. Within our data set, all but two coincident peak fell during a warning period. Further, upon discussion with the manager of the CPP program, these two mistakes are attributed to human error rather than an unpredicted coincident peak.

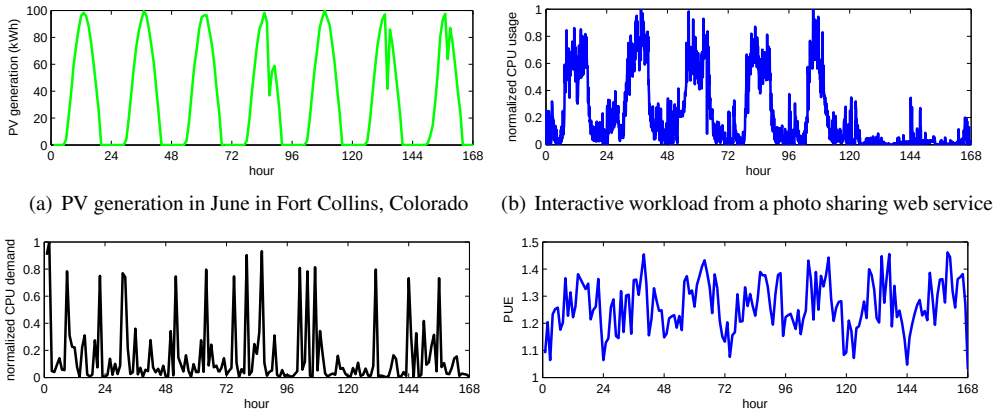
Figure 1(c) shows the number of warnings on the time of the day. Given that the warnings are well correlated with the coincident peak shown in Figure 1(a), it is important to understand their frequency and timing. Unsurprisingly, the announcement of warnings has strong diurnal pattern similar to that of the coincident peak: most warnings happen between 2pm and 10pm. The seasonal pattern is also similar to that of the coincident peak: winter months have warnings later in the day than summer months, and the time range in which most warnings occur is narrower during winter months. Additionally, summer months have significantly more warnings than winter month do (14.5 warnings per month in summer compared to 10 in winter). The number of warnings over the week is shown in Figure 1(d). Similar to that of the coincident peak shown in Figure 1(b), the warnings have a strong weekly pattern: few warnings happen during the weekends, and the number of warnings decreases throughout the weekdays.

Some other interesting phenomena are shown in Figure 2. In particular, the frequency of warnings decreases during the month, the length of consecutive warnings tends to be 2-4 hours. the number of warnings in a month varies from 3 to 22, and the number of days with warnings during a month tends to be less than seven.



(a) Frequency of warnings during a month (b) Length of consecutive warnings (c) Number of warnings per month (d) Number of days with warnings per month

Figure 2. Overview of warning occurrences showing (a) daily frequency, (b) length, and (c)-(d) monthly frequency.



(a) PV generation in June in Fort Collins, Colorado (b) Interactive workload from a photo sharing web service

(c) Facebook Hadoop workload (d) Google data centers PUE

Figure 3. One week traces for (a) PV generation, (b) non-flexible workload demand, (c) flexible workload demand, and (d) cooling efficiency.

3. Modeling

The core of our approach for developing data center demand response algorithms is an energy expenditure model for a data center participating in a CPP program. We introduce our model for data center energy costs in this section. It builds on the model used by Liu et al. in [23], which is in turn related to the models used in [24, 25, 26, 27, 12, 28, 29, 30]. The key change we make to [23] is to incorporate charges from CPP, workload demand and renewable generation prediction errors into the objective function of the optimization. This is a simple modeling change, but one that creates significant algorithmic challenges (see Section 4 for more details).

Our cost model is made up of models characterizing the power supply and power demands of a data center. On the power supply side, we model a power micro-grid consisting of the public grid, local backup power generation, and/or a renewable energy supply. On the power demand side, we consider both non-flexible interactive workloads and flexible batch-style workloads in the data centers. Further, we consider a cooling model that allows for a mixture of different cooling methods, e.g., "free" outside air cooling and traditional mechanical chiller cooling.

Throughout, we consider a discrete-time model whose time slot matches the time scale at which the capacity provisioning and scheduling decisions can be updated. There is a (possibly long) planning horizon that we are interested in, $\{1, 2, ..., T\}$. In practice, T could be a day and a time slot length could be 1 hour.

3.1. Power Supply Model

The electricity cost from the grid includes three non-constant components as described in Section 2, denote by $p(t)$ the usage price, p_p the (customer) peak price, and p_{cp} the coincident peak price. We assume all prices are positive without loss of generality.

Most data centers are equipped with local power generators as backup, e.g., diesel or natural gas powered generators. These generators are primarily intended to provide power in the event of a power failure; however they can be valuable for data center demand response, e.g., shedding peak load by powering the data center with local generation. Typically, the costs of operating these generators are dominated by the cost of fuel, e.g., diesel or natural gas. Note that the effective output of such generators can often be adjusted. In many cases the backup generation is provided by multiple generators which can be operated independently [31], and in other cases the generators themselves can be adjusted continuously, e.g., in the case of a GE gas engine [32].

To model such local generators, we assume that the generator has the capacity to power the whole data center, which is quite common in industry [31], i.e., the total capacity of local generators $C_g = C$, where C is the total data center power capacity. We denote the cost in dollar of generating 1kWh power using backup generator by p_g. Finally, we denote the generation provided by the local generator at time t by $g(t)$.

In addition to local backup generators, data centers today increasingly have some form of local renewable energy available such as PV [33]. The effective output of this type of generation is not controllable and is often closely tied to external conditions (e.g., wind speed and solar irradiance). Figure 3(a) shows the power generated from a 100kW PV installation in June in Fort Collins, Colorado. The fluctuation and variability present a significant challenge for data center management. In this paper, we consider both data centers with and without local renewable generation. To model this, we use $r(t)$ to denote the actual renewable energy available to the data center at time t and use $\hat{r}(t)$ for the predicted generation. We denote $r(t) = (1 + \hat{\epsilon}_r)\hat{r}(t)$, where $\hat{\epsilon}_r$ is the prediction error. We assume unbiased prediction $\mathbb{E}[\hat{\epsilon}_r] = 0$ and denote the variance $\mathbb{V}[\hat{\epsilon}_r]$ by σ_r^2, which can be obtained from historic data. These are standard assumptions in statistics. Let $\hat{\xi}_r$ denote the distribution of $\hat{\epsilon}_r$. In the model, we ignore all fixed costs associated with local generation, e,g., capital expenditure and renewable operational and maintenance cost.

3.2. Power Demand Model

The power demand model is derived from models of the workload and the cooling demands of the data center.

Workload model. Most data centers support a range of IT workloads, including both non-flexible interactive applications that run 24x7 (such as Internet services, online gaming) and delay tolerant, flexible batch-style applications (e.g., scientific applications, financial analysis, and image processing). Flexible workloads can be scheduled to run anytime as long as the jobs finish before their deadlines. These deadlines are much more flexible (several hours to multiple days) than that of interactive workload. The prevalence of flexible workloads provides opportunities for providing demand response via workload shifting/shaping.

We assume that there are I interactive workloads. For interactive workload i, the arrival rate at time t is $\lambda_i(t)$. Then based on the service rate and the target performance metrics (e.g., average delay, or 95th percentile delay) specified in SLAs, we can obtain the IT capacity required to allocate to each interactive workload i at time t, denoted by $a_i(t)$. Here $a_i(t)$ can be derived from either analytic performance models, e.g., [34], or system measurements as a function of

$\lambda_i(t)$ because performance metrics generally improve as the capacity allocated to the workload increases, hence there is a sharp threshold. Interactive workloads are typically characterized by highly variable diurnal patterns. Figure 3(b) shows an example from a 7-day normalized CPU usage trace for a popular photo sharing and storage web service which has more than 85 million registered users in 22 countries.

Flexible batch jobs are more difficult to characterize since they typically correspond to internal workloads and are thus harder to attain accurate traces for. Figure 3(c) shows an example from a 7-day normalized CPU demand trace generated using arrival and job information about Facebook Hadoop workload [35, 36]. We assume there are J classes of batch jobs. Class j jobs have total demand B_j, maximum parallelization MP_j, starting time S_j and deadline E_j. Let $b_j(t)$ denote the amount of capacity allocated to class j jobs at time t. We have $0 \leq b_j(t) \leq MP_j, \forall t$ and $\sum_{t\in[S_j,E_j]} b_j(t) = B_j$.

Given the above models for interactive and batch jobs, the total IT demand at time t is given by

$$d_{IT}(t) = \sum_{i=1}^{I} a_i(t) + \sum_{j=1}^{J} b_j(t). \tag{1}$$

The total IT capacity in units of kWh is D, so $0 \leq d_{IT}(t) \leq D, \forall t$. Since our focus is on energy costs, we interpret $d_{IT}(t)$, $a_i(t)$, and $b_j(t)$ as being the energy necessary to serve the demand, and thus in units of kWh.

Cooling model. In addition to the power demands of the workload itself, the cooling facilities of data centers can contribute a significant portion of the energy costs. Cooling power demand depends fundamentally on the IT power demand, and so is derived from IT power demand through cooling models, e.g., [37, 38]. Here, we assume the cooling power associated with IT demand d_{IT}, $c(d_{IT})$, is a convex function of d_{IT}. One simple but widely used model is Power Usage Effectiveness (PUE) as follows: $c(d(t)) = (PUE(t) - 1) * d(t)$. Note that $PUE(t)$ is the PUE at time t, and varies over time depending on environmental conditions, e.g., the outside air temperature. Figure 3(d) shows one week from a trace of the average PUE of Google data centers. More complex models of the cooling cost have also been derived in the literature, e.g., [23, 37, 38].

Total power demand. The total power demand is denoted by

$$d(t) = d_{IT}(t) + c(d_{IT}(t)). \tag{2}$$

We use $\hat{d}(t)$to denote the predicted demand. We denote $d(t) = (1 + \hat{\epsilon}_d)\hat{d}(t)$, where $\hat{\epsilon}_d$ is used to stand for the prediction error. Again, we assume $\mathbb{E}[\hat{\epsilon}_d] = 0$ and denote $\mathbb{V}[\hat{\epsilon}_d]$ by σ_d^2, which can be obtained from historic data. Let $\hat{\xi}_d$ denote the distribution of $\hat{\epsilon}_d$.

3.3. Total data center costs

Using the above models for the power supply and power demand at a data center, we can now model the operational energy cost of a data center, which our data center demand response algorithms seek to minimize. In particular, they take the power supply cost parameters, including the grid power pricing and fuel cost, as well as the workload demand and SLAs information, as input and seek to provide an near-optimal workload schedule and a local power generation plan given uncertainties about workload demand and renewable generation. This planning problem can be formulated

as the following constrained convex optimization problem given t_{cp}.

$$\min_{\mathbf{b},\mathbf{g}} \quad \sum_{t=1}^{T} p(t)e(t) + p_p \mathbf{max}_t e(t) + p_{cp} e(t_{cp}) + p_g \sum_{t=1}^{T} g(t) \tag{3a}$$

$$\text{s.t.} \quad e(t) \equiv (d(t) - r(t) - g(t))^+ \leq C, \quad \forall t \tag{3b}$$

$$\sum_{t \in [S_j, E_j]} b_j(t) = B_j, \quad \forall j \tag{3c}$$

$$0 \leq b_j(t) \leq MP_j, \quad \forall j, \forall t \tag{3d}$$

$$0 \leq d_{IT}(t) \leq D, \quad \forall t \tag{3e}$$

$$0 \leq g(t) \leq C_g. \quad \forall t \tag{3f}$$

In the above optimization, the objective (3a) captures the operational energy cost of a data center, including the electricity charge by the utility and the fuel cost of using local power generation. The first three terms describe grid power usage charge, peak demand charge, and coincident peak charge, respectively. The fuel cost of the local power generator is specified in the last term. Further, the first constraint (3b) defines $e(t)$ to be the power consumption from the grid at time t, which depends on the IT demand $d_{IT}(t)$ defined in (1) and therefore further depends on batch job scheduling $b_j(t)$, the cooling demand, the availability of renewable energy, and the use of the local backup generator. Constraint (3c) requires all jobs to be completed. Constraint (3d) limits the parallelism of the batch jobs. Constraint (3e) limits the demand served in each time slot by the IT capacity of the data center. The final constraint (3f) limits the capacity of the local generation.

4. Algorithms

We now present our algorithms for workload and generation planning in data centers that participate in CPP programs. In particular, our starting point is the data center optimization problem introduced in (3a) in the previous section, and our goal is to design algorithms for optimally combining local generation and workload shifting in order to minimize the operational energy cost. More specifically, the algorithmic problem we approach is as follows. We assume that the planning horizon being considered is one day and that the workload, prices, cooling efficiency, and renewable availability can be predicted with reasonable accuracy in this horizon, but that the planner does not know when the coincident peak and the corresponding warnings will occur. The algorithmic goal is thus to generate a plan that minimizes cost despite this unknown information and prediction errors. Since the costs associated with the coincident peak can be a large fraction of the data center electricity bill, this lack of information is a significant challenge for planning. As we have already discussed, designing for this uncertainty about the coincident peak is fundamentally different than designing for prediction errors on factors such as workload demand or renewable generation since inaccuracies in the prediction of the coincident peak and the corresponding warnings have a discontinuous threshold effect on the realized cost. As a result, even small prediction errors can result in significantly increased costs. Such effects are well-known to make the design of online algorithms difficult.

We consider two approaches for handling uncertainty about the coincident peak. The first approach we follow is to estimate when the coincident peak and the corresponding warnings will occur. Using the estimated likelihood

of a warning and/or coincident peak during each hour, we can formulate a convex optimization problem to minimize the *expected* cost in the planning horizon. The second approach we follow is to formulate a robust optimization that seeks to minimize the *worst case* cost given adversarial placement of warnings and the coincident peak. Note that throughout this paper we restrict our attention to algorithms that do "non-adaptive" workload shifting, i.e., algorithms that plan workload shifting once at the beginning of the horizon and then do not adjust the plan during the horizon in order to make them more easily adoptable. However, we do allow local generation to be turned on adaptively when warnings are received. This restriction is motivated by industry practice today – adaptive workload shifting for demand response is nearly non-existent, but data centers that actively participate in demand response programs do adjust local generation when warnings are received. This restriction can easily be relaxed in what follows.[1] However, the fact that our analytic results provide guarantees for non-adaptive workload planning means they are *stronger*. Further, our numerical experiments studying the improvements from adaptive workload planning (omitted due to space restrictions) highlight that the benefit of such adaptivity is not large. This can be seen already in our results since the gap between the costs of our non-adaptive algorithms and the cost of the offline optimal is small.

4.1. Expected cost optimization

The starting point for our algorithms is the data center optimization in (3a). In this section, our goal is to plan workload allocation and local generation in order to minimize the expected cost of the data center given estimates from historical data about when the warnings and the coincident peak will occur. In particular, our approach uses historical data about when warnings will occur in order to estimate the likelihood that time slot t will be a warning. We denote the estimate at time t by $\hat{w}(t)$, and the full estimator by $\hat{\mathsf{W}}$.

Since the data center has local backup generation, it can provide demand response even without using adaptive workload shifting by turning on the backup generator when warnings are received from the utility. Today, those data centers that actively participate in demand response programs typically use this approach. The reason is that the cost of local generation is typically significantly less than the coincident peak price, and the number of warnings per month is small enough to ensure that it is cost efficient to always turn on generation whenever warnings are given. Of course, there are drawbacks to using local generation, since it is typically provided by diesel generators, which often have very high emissions and costs [39, 4]. Thus, it is important to do workload shifting in a manner that minimizes the use of local generation, if possible.

Before stating the algorithm formally, let briefly discuss its structure. Using the estimates of warning occurrences, workload demand and renewable generation, we first solve a stochastic optimization (given in Algorithm 1 below) to obtain a workload schedule $b(t)$ and local generator usage plan $g_1(t)$. Then, in runtime, when the prediction error is harmful, i.e., when

$$\mathbf{min}\{e(t), \epsilon_d \hat{d}(t) - \epsilon_r \hat{r}(t)\} > 0, \tag{4}$$

use the backup generator to remove this effect, i.e., use generation $g_\epsilon(t) = \mathbf{max}\{0, \mathbf{min}\{(e(t), \epsilon_d \hat{d}(t) - \epsilon_r \hat{r}(t)\}\}$.[2] Additionally, if a warning occurs, turn on the local generator to reduce the demand from the grid to zero, which we denote

[1] If it is relaxed, replanning after warnings occur can be beneficial. Interestingly, such replanning could have similar and only slightly better performance in the worst case. We omit the results due to space constraints.

by $g_2(t) = e(t) - g_\epsilon(t)$ when t is a warning period in order to ensure that the coincident peak payment is zero. (Recall that the coincident peak happens within a warning period with near certainty.) The total local generation used is thus $g(t) = g_1(t) + g_\epsilon(t) + g_2(t), \forall t$. More formally, to write the objective function used for the first step of planning we first need to estimate $g_2(t)$, which can be done as follows:

$$g_2(t) = \begin{cases} e(t) - g_\epsilon(t) & \text{if } t \text{ is a warning hour} \\ 0 & \text{otherwise} \end{cases}$$

This is feasible since in practice the generator has the capacity to power the whole data center [31], i.e., $C_g = C$.

We can now formally define the planning algorithm for expected cost minimization. Define $\hat{e}(t) \equiv \left(\hat{d}(t) - \hat{r}(t) - g_1(t)\right)^+$ as the predicted power demand from utility at time t, and $\sigma \equiv \mathbf{max}\{\sigma_d, \sigma_r\}$ as a upper bound of normalized variance of the power demand from utility.

Algorithm 1. *Estimate $\hat{w}(t)$ for all t in the planning period. Then, solve the following convex optimization:*

$$\min_{\mathbf{b},\mathbf{g}_1} \quad \sum_{t=1}^{T} \left((1-\hat{w}(t))p(t) + \hat{w}(t)p_g\right)\hat{e}(t) + p_p \mathbf{max}_t \hat{e}(t) + p_g \sum_{t=1}^{T} g_1(t) \tag{5a}$$

$$\text{s.t.} \quad \hat{e}(t) \equiv \left(\hat{d}(t) - \hat{r}(t) - g_1(t)\right)^+ \le C, \quad \forall t \tag{5b}$$

$$\sum_{t \in [S_j, E_j]} b_j(t) = B_j, \quad \forall j \tag{5c}$$

$$0 \le b_j(t) \le MP_j, \quad \forall j, \forall t \tag{5d}$$

$$0 \le \hat{d}(t) \le D, \quad \forall t \tag{5e}$$

$$0 \le g_1(t) \le C_g. \quad \forall t \tag{5f}$$

During operation, if the prediction error has negative effect satisfying (4)*, use backup generation to remove the error.*[2] *If a warning is received, use the local generator to reduce the power usage from the grid to zero until the warning period ends.*

Of course there are many approaches for estimating $\hat{w}(t)$ in practice. In this paper, we do this using the historical data summarized in Section 2. Since our data is rich, and the occurrence of the warnings is fairly stationary, this estimator is accurate enough to achieve good performance, as we show in Section 5. Of course, in practice predictions could likely be improved by incorporating information such as weather predictions.

It is clear that the performance of Algorithm 1 is highly dependent on the accuracy of predictions, thus it is important to characterize this dependence. To accomplish this, denote the objective function in (3a) by $f(\mathbf{b}, \mathbf{g})$. Then the expected cost of Algorithm 1 is $\mathbb{E}_{\hat{\xi}_d, \hat{\xi}_r, \hat{W}}\left[f(\mathbf{b}^s, \mathbf{g}^s)\right]$. We compare this cost to the expected cost of oracle-like offline algorithm that knows workload demand and renewable generation perfectly, which we denote by $\mathbb{E}_{\hat{\xi}_d, \hat{\xi}_r, \hat{W}}\left[f(\mathbf{b}^*, \mathbf{g}^*)\right]$. To characterize the performance of the algorithm we use the *competitive ratio*, which is defined as the ratio of the

[2] Note that, in practice, one would not want to use generation to correct for *all* prediction errors, such a correction would only be done if the prediction error was extreme. However, for analytic simplification, we assume that all prediction errors are erased in this manner and evaluate the resulting cost. Our simulations results in Section 5 use the generator only to correct for extreme prediction errors.

cost of a given algorithm to the cost of the offline optimal algorithm. The following theorem (proven in Appendix A) shows that the cost of the online algorithm is not too much larger than optimal as long as predictions are accurate.

Theorem 1. *Given that the standard deviation of prediction errors for the workload and renewable generation are bounded by σ and the distribution of coincident peak warnings is known precisely, Algorithm 1 has a competitive ratio of* $1 + B\sigma$, *where* $B = \frac{p_g \Sigma_t (\hat{d}^s(t)+\hat{r}(t))}{2\mathbb{E}_{\varepsilon_{\mathbf{d}}}[f^*(\mathbf{e}^*,\mathbf{g}^*)]} + \frac{p_g \Sigma_t (\hat{d}^*(t)+\hat{r}(t))}{2\mathbb{E}_{\varepsilon_{\mathbf{d}}}[f^*(\mathbf{e}^*,\mathbf{g}^*)]}$. *That is,* $\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r,\hat{W}}\left[f(\mathbf{b}^s,\mathbf{g}^s)\right] / \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r,\hat{W}}\left[f(\mathbf{b}^*,\mathbf{g}^*)\right] \leq 1 + B\sigma$.

It is worth noting that it is rare for the impact of prediction error on a data center planning algorithm to be quantified analytically, nearly all prior work either does not study the impact of prediction errors, or studies their impact via simulation only. Additionally, it is important to point out that Theorem 1 does not make any distributional assumption on the prediction errors other than bounded variance. The key observation provided by Theorem 1 is that the competitive ratio is a linear function of prediction standard deviation, which implies when prediction errors decrease to 0, this competitive ratio also decreases to 1. Thus, the algorithm is fairly robust to prediction errors. Our trace-based simulations in Section 5 corroborate this conclusion.

4.2. Robust optimization

While performing well for expected cost is a natural goal, the algorithm we have discussed above depends on the accuracy of estimators of the occurrence of the coincident peak or warning periods. In this section, we focus on providing algorithms that maintain worst-case guarantees regardless of prediction accuracy, i.e., that minimize the worst case cost. To characterize the performance of the algorithm we again use the *competitive ratio*. In our setting, we consider the cost only during one planning period. Thus, the difference in information between the offline algorithm and our algorithm is knowledge of when the warnings will occur, exact workload demand and renewable generation. We do assume that the online algorithm has an upper bound on the number of warnings that may occur.

In order to minimize the worst case cost, the natural approach is to increase the penalty on the peak period. This follows because, if an adversary seeks to maximize the cost of an algorithm, it should place warnings on the periods where the algorithm uses the most energy. This observation leads us to the following algorithm:

Algorithm 2. *Consider an upper bound on the number of warning periods $\bar{W}$. Solve the following convex optimization*

$$\min_{\mathbf{b},\mathbf{g}_1} \quad \sum_{t=1}^{T} p(t)\hat{e}(t) + \left(p_p + \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\right)\mathbf{max}_t \hat{e}(t) + p_g \sum_{t=1}^{T} g_1(t) \tag{6a}$$

$$\text{s.t.} \quad \hat{e}(t) \equiv \left(\hat{d}(t) - \hat{r}(t) - g_1(t)\right)^+ \leq C, \ \ \forall t \tag{6b}$$

$$\sum_{t\in[S_j,E_j]} b_j(t) = B_j, \ \ \forall j \tag{6c}$$

$$0 \leq b_j(t) \leq MP_j, \ \ \forall j, \forall t \tag{6d}$$

$$0 \leq \hat{d}(t) \leq D, \ \ \forall t \tag{6e}$$

$$0 \leq g_1(t) \leq C_g. \ \ \forall t \tag{6f}$$

During operation, if the prediction error has negative effect satisfying (4)*, use backup generation to remove the error.*[2] *If a warning is received, use the local generator to reduce the power usage from the grid to zero until the warning period ends.*

This algorithm represents a seemingly easy change to the original data center optimization in (3a); however the subtle differences are enough to ensure that it provide a very strong worst case cost guarantee. In particular, it provides the minimal competitive ratio achievable.

Theorem 2. *Given that the standard deviation of prediction errors for the workload and renewable generation are bounded by σ, Algorithm 2 has a competitive ratio of*

$$1 + B\sigma + \frac{\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)}{T\mathbf{min}_t p(t)/PMR^* + p_p} \leq 1 + B\sigma + \frac{\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)}{p_p},$$

where $B = \frac{p_g \Sigma_t(\hat{d}^w(t)+\hat{r}(t))}{2\mathbb{E}_{\varepsilon_\mathbf{d}}[f^*(\mathbf{e}^*,\mathbf{g}^*)]} + \frac{p_g \Sigma_t(\hat{d}^*(t)+\hat{r}(t))}{2\mathbb{E}_{\varepsilon_\mathbf{d}}[f^*(\mathbf{e}^*,\mathbf{g}^*)]}$. Further, if $\bar{W} = |W|$ then there is a lower bound $1 + \frac{\bar{W}(p_g - \mathbf{min}_t p(t))}{T\mathbf{min}_t p(t)/PMR^* + p_p}$ on the competitive ratio achievable under any online algorithm, even one with exact predictions of workloads and renewable generation.

The key contrast between Theorem 2 and Theorem 1 is that Theorem 1 assumes that the distribution of coincident peak warnings is known precisely, while Theorem 2 provides a bound even when the coincident peak warnings are adversarial. As such, it is not surprising that the competitive ratio is larger in Theorem 2. However, note that the competitive ratio of Algorithm 1 in the context of Theorem 2 can be easily shown to be unbounded, and so one should not think of Theorem 1 as a stronger bound than Theorem 2.

Interestingly, the form of Theorem 2 parallels Theorem 1, except with an additional term in competitive ratio. Thus, again the competitive ratio grows linearly with the variance of the prediction error. Additionally, note that when $\sigma = 0$, the competitive ratio matches the lower bound, which highlights that the additional term in Theorem 2 is tight. Further, since the additional term is defined in terms of the relative prices of local generation and the peak, it is easy to understand its impact in practice. In practice, p_g is less than \$0.3/kWh [40] and the number of warning hours is roughly between 3 and 22, with an average of 12 warning hours per month. So, this term is typically less than 1, which highlights that the worst-case bound on Algorithm 2 nearly matches the bound on Algorithm 1 in the case where the coincident peak warning distribution is known.

Note that, if there is no local generator, then we can derive a similar result to Theorem 2, where $\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)$ is replaced by p_{cp}. The comparison of these results highlights the cost savings provided by using a local backup generator. Since the data center does not know the exact number of warnings for a particular month, whether or not using local generation is beneficial depends on the predicted bound on the number of warnings per month. If it is smaller than $\left\lfloor \frac{p_{cp}}{p_g - \mathbf{min}_t p(t)} \right\rfloor$ (25 in winter and 36 in summer for 2012 in the utility scheme shown in Table 1 with high local generation cost), it should use local generation. This highlights that if a utility wishes to incentivize the data center to use local generation to relieve its pressure, then it should not send too many warnings.

4.3. Implementation considerations

Over the past decade there has been significant effort to address data center energy challenges via workload management. Most of these efforts focus on improving the energy efficiency and achieving energy proportionality of data centers via workload consolidation and dynamic capacity provisioning, e.g., [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Recently, such work has begun to explore topics such as shifting (temporal) or migrating (spatial) workloads to better use renewable energy sources [41, 25, 28, 42, 43, 29, 44, 45].

The algorithms presented in this section are both optimization-based approaches for temporal workload management and, as such, build on this literature. In particular, optimization based approaches have received significant attention in recent years, and have been shown to transition easily to large scale implementations, e.g., [23, 10, 5]. In this paper, we evaluate the algorithms presented above via both worst-case analysis and trace-based simulations. However, for completeness we comment briefly here on the important considerations for implementation of these designs. For more details, the reader should consult [23, 10, 5]. Implementation considerations typically fall into two categories: (i) obtaining accurate predictions of workload, renewable generation, costs, etc.; (ii) implementing the plan generated by the algorithm. Each of these challenges has been well studied by prior literature, and we only provide a brief description of each in the following.

Predictions. Our algorithms exploit the statistical properties of the coincident peak as well as predictions of IT demand, cooling costs, renewable generation, etc. Historical data about the coincident peak is generally available, for large industrial consumers, from the utilities operating demand response programs. In practice, coincident peak predictions can also be improved using factors such as the weather. Other parameters needed by our algorithm are also fairly predictable. For example, in a data center with a renewable supply such as a solar PV system, our planning algorithms need the predicted renewable generation as input. This can be done in many ways, e.g., [46, 23, 44] and a ballpark approximation is often sufficient for planning purposes. Similarly, IT demands typically exhibit clear short-term and long-term patterns. To predict the resource demand for interactive applications, we can first perform a periodicity analysis of the historical workload traces to reveal the length of a pattern or a sequence of patterns that appear periodically via Fast Fourier Transform (FFT). An auto-regressive model can then be created and used to predict the future demand of interactive workloads. For example, this approach was followed by [23]. The total resource demand (e.g., CPU hours) of batch jobs can be obtained from users or from historical data or through offline benchmarking [47]. Like supply prediction, a ballpark approximation is typically good enough. Finally, there are many approaches for deriving cooling power from IT demand, for example the models in [37, 23].

Execution. Given the predictions for the coincident peak, IT demand, cooling costs, renewable generation, etc., our proposed algorithms proceed by solving an optimization problem to determine a plan. Since the optimization problems used are convex and in simple form, they can be solved efficiently. Given the resulting plan, the remaining work is to implement the actual workload placement and consolidation on physical servers. This can be done using packing algorithms, e.g., simple techniques such as Best Fit Decreasing (BFD) or more advanced algorithms such as [48]. Finally, the execution of the plan can be done by a runtime workload generator, which schedules flexible workload and allocates CPU resources according to the plan. This can be easily implemented in virtualized environments. For example, a KVM or Xen hypervisor enables the creation of virtual machines hosting batch jobs; the adjustment of the resource allocation (e.g., CPU shares or number of virtual CPUs) at each virtual machined; and the migration and consolidation of virtual machines. An example using this approach is [23]. Further, [5] provides more concrete details of implementing the plan in the field. These suggest that the benefits from our algorithms are attainable in real system, and we will focus on numerical simulations in the following section.

5. Case study

To this point we have introduced two algorithms for managing workload shifting and local generation in a data center participating in a CPP program. We have also provided analytic guarantees on these algorithms. However, to get a better picture of the cost savings such algorithms can provide in practical settings, it is important to evaluate the algorithms using real data, which is the goal of this section. We use numerical simulations fed by real traces for workloads, cooling efficiency, electricity pricing, coincident peak, etc., in order to contrast the energy costs and emissions under our algorithms with those under current practice.

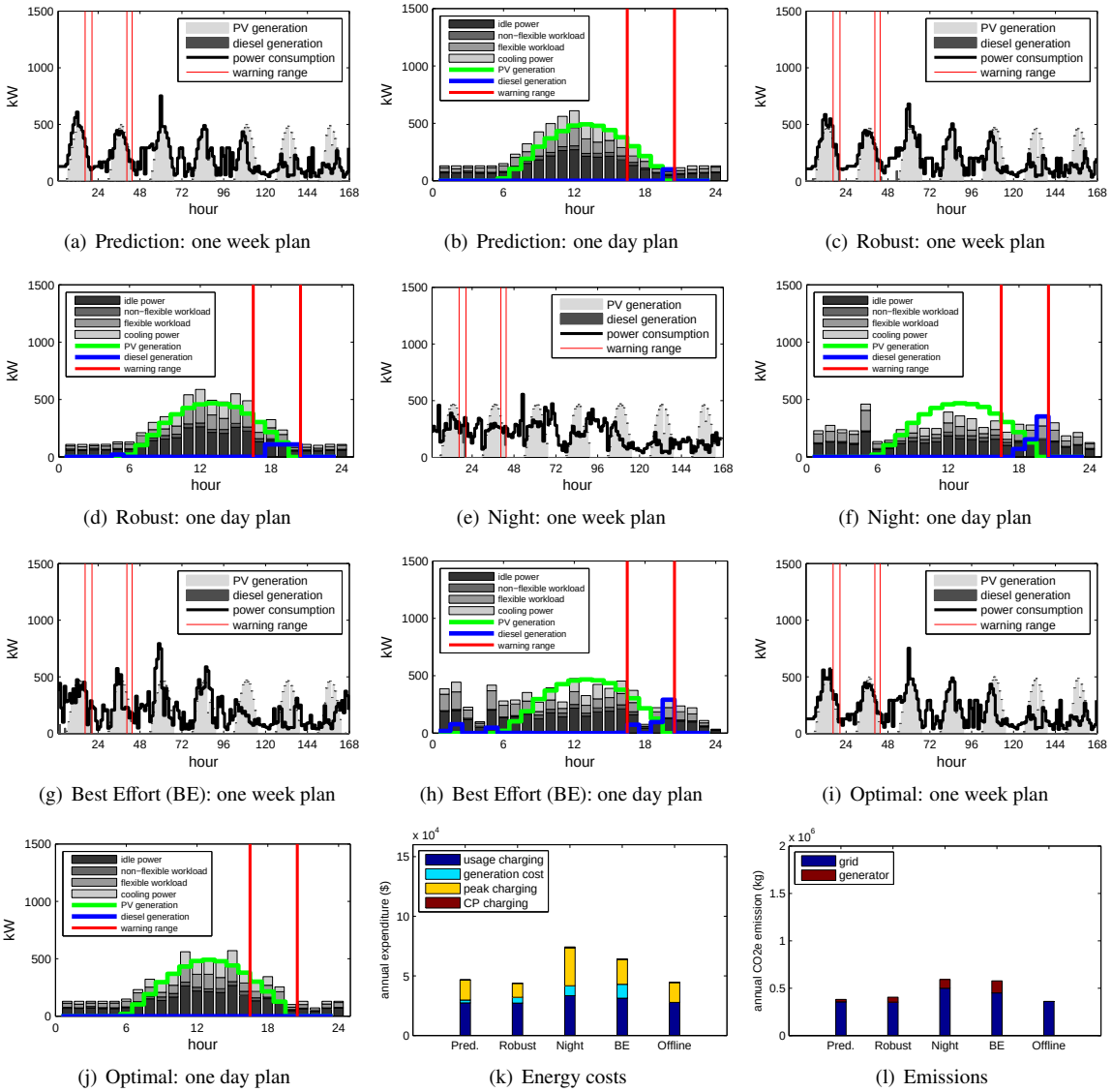
5.1. Experimental setup

Workload and cost settings. To define the workload for the data center we use traces from real data centers for interactive IT workload, batch jobs, and cooling data. The interactive workload trace is from a popular web service application with more than 85 million registered users in 22 countries (see Figure 3(b)). The trace contains average CPU utilization and memory usage as recorded every 5 minutes. The peak-to-mean ratio of the interactive workload is about 4. The batch job information comes from a Facebook Hadoop trace (see Figure 3(c)). The total demand ratio between the interactive workload and batch jobs is 1:1.6. This ratio can vary widely across data centers, and our previous work studied its impacts [23]. The deadlines for the batch jobs are set so that the lifespan is 4 times the time necessary to complete the jobs when they are run at their maximum parallelization. The maximum parallelization is set to the total IT capacity divided by the mean job submission rate. The time varying cooling efficiency trace is derived from Google data center data and the PUE (see Figure 3(d)) is between 1.1 and 1.5. The prediction error of workload and cooling power demand has a standard deviation of 10% from our simple prediction algorithm. The total IT capacity is set to 3500 servers (700kW). Server idle power is 100W and peak power is 200W. The energy related costs are determined from the Fort Collins Utilities data described in Section 2. The prices are chosen to be the 2011 rates in Table 1. The local power generation of the data center is set as follows. In different settings the data center may have both a local diesel generator and a local PV installation[3]. When a diesel generator is present, we assume it has the capacity to power the full data center, which is set to be 1000kW. The cost of generation is set at \$0.3/kWh [40] for conservative estimates. The emissions are set to be 3.288kg CO_2 equivalent per kWh [39]. The emission of grid power is set to be 0.586kg CO_2 equivalent per kWh [40]. The PV capacity is set to be 700kW and the prediction error of PV generation has standard deviation 15% from our prediction algorithm.

Comparison baselines. In our experiments, our goal is to evaluate the performance of the algorithms presented in Section 4. We consider a planning period that is 24-hours starting at midnight. The planner determines workload shifting and local generation usage at an hourly level, i.e., the amount of capacity allocated to each batch job and the amount of power generated by the local diesel generator at each time slot. The length of each time slot is one hour.

In this context, we compare the energy costs and emissions of the algorithms presented in Section 4 with two baselines, which are meant to model industry standard practice today. In our study, Algorithm 1 is termed "*Prediction (Pred)*", which utilizes predictions about the coincident peak warnings to minimize the expected cost. Similarly, Algorithm 2 optimizes the worst-case cost, and are termed "*Robust*". The baseline algorithms are "*Night*", "*Best*

[3] we have more results about other combinations, but omit due to space constraint.



(a) Prediction: one week plan (b) Prediction: one day plan (c) Robust: one week plan

(d) Robust: one day plan (e) Night: one week plan (f) Night: one day plan

(g) Best Effort (BE): one week plan (h) Best Effort (BE): one day plan (i) Optimal: one week plan

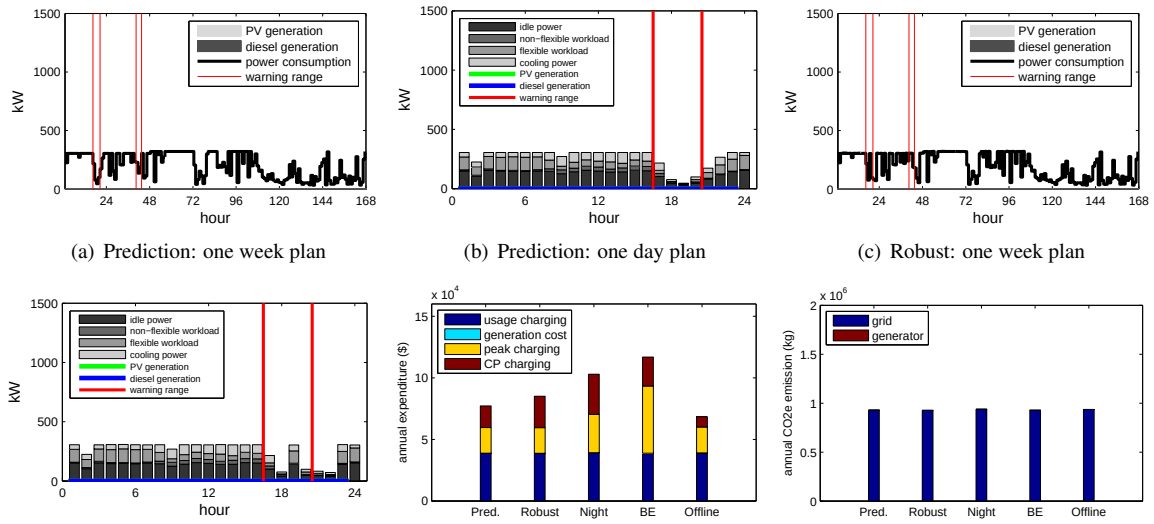
(j) Optimal: one day plan (k) Energy costs (l) Emissions

Figure 4. Comparison of energy costs and emissions for a data center with a local PV installation and a local diesel generator. (a)-(j) show the plans computed by our algorithms and the baselines.

Effort (BE)", and "*Optimal*". *Night* and *Best Effort* are meant to mimic typical industry heuristics, while *Optimal* is the offline optimal plan given knowledge of when the coincident peak will occur, exact workload demand and renewable generation. *Best Effort* finishes jobs in a first-come-first-serve manner as fast as possible. *Night* tries to run jobs during night if possible and otherwise run these jobs with a constant rate to finish them before their deadlines.

5.2. Experimental results

In our experimental results, we seek to explore the following issues: (i) How much cost and emission savings can our algorithms achieve? How close to optimal are our algorithms on real workloads? (ii) What are the relative benefits of local generation and workload shifting and a mixture of both with respect to cost and emission reductions? (iii)



(a) Prediction: one week plan (b) Prediction: one day plan (c) Robust: one week plan

(d) Robust: one day plan (e) Energy costs (f) Emissions

Figure 5. Comparison of energy costs and emissions for a data center without local generation or PV generation. (a)-(d) show the plans computed by our algorithms.

What is the impact of errors in predictions of the coincident peak and the corresponding warnings?

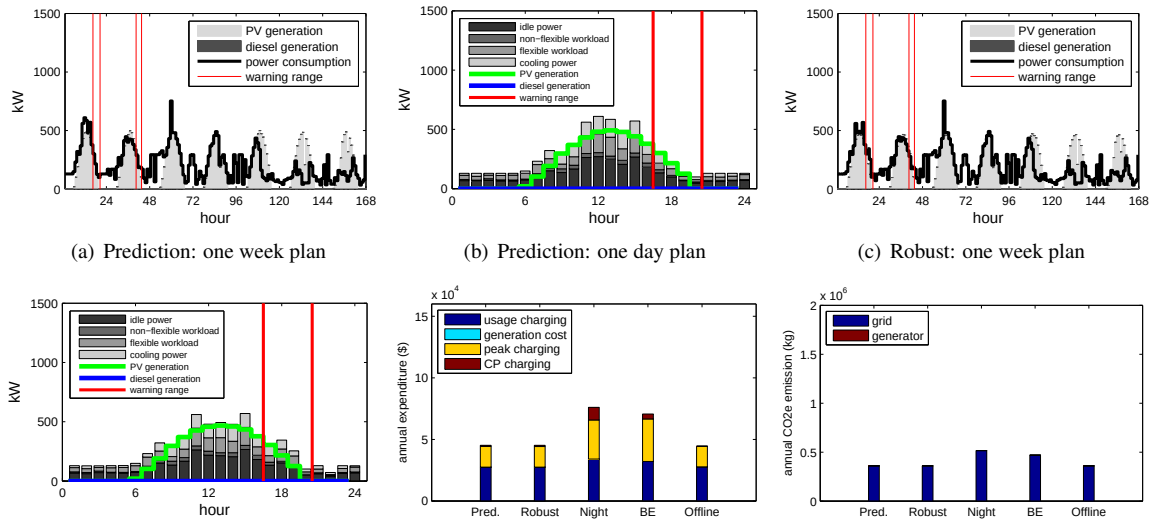
5.2.1. Cost savings and emissions reductions

We start with the key question for the paper: how much cost and emission savings do our algorithms provide? Figure 4 shows our main experimental results comparing our algorithms with baselines. The weekly power profile for the first week of June 2011 is shown in the first plot for each algorithm, including power consumption, PV generation and diesel generation, and coincident peak warnings. The detailed daily power breakdown for the first Monday in June 2011 is shown in the second plot for each algorithm, including idle power, power consumed by serving flexible workload and non-flexible workload, cooling power, local generation and warnings. Further, the last two plots includes a cost comparison and an emissions comparison for over *one year* of operation, including usage costs, peak costs, CP costs, local generation costs, and emissions from both the grid power and local generation used.

As shown in the figure, our algorithms provide 40% savings compared to *Night* and *Best Effort*. Specifically, *Prediction* reshapes the flexible workload to prevent using the time slots that are likely to be warning periods or the coincident peak as shown in Figures 4 (a) and (b), while *Robust* tries to make the grid power usage as flat as possible as shown in Figures 4 (c) and (d). Both algorithms try to fully utilize PV generation. In contrast, *Night* and *Best Effort* do not consider the warnings, the coincident peak, or renewable generation. Therefore, they have significantly higher coincident peak charges and local generation costs (*Night* has higher cost here because it wastes even more renewable generation). Since the warning and coincident peak predictions are quite accurate, *Prediction* works better than *Robust* and similar to *Optimal*.

5.2.2. Local generation versus workload shifting

A second important goal of this paper is to understand the relative benefits of local generation planning and workload shifting for data centers participating in CPP programs. Though our algorithms have focused on the case



(a) Prediction: one week plan (b) Prediction: one day plan (c) Robust: one week plan

(d) Robust: one day plan (e) Energy costs (f) Emissions

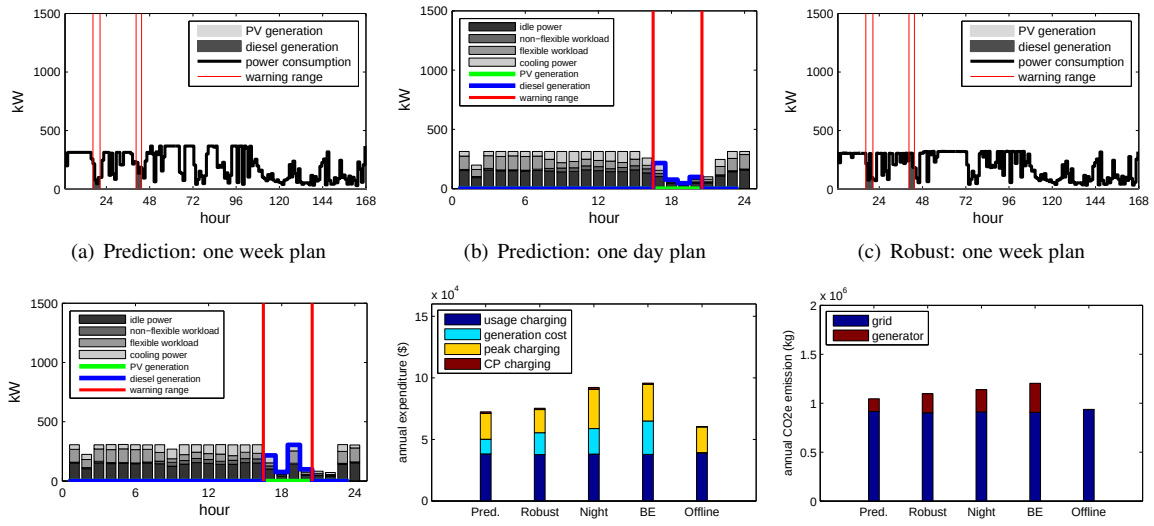
Figure 6. Comparison of energy costs and emissions for a data center with a local PV installation, but without local generation. (a)-(d) show the plans computed by our algorithms.

of local generation, they can be easily adjusted to the case where there is no local generator. In fact, similar analytic results hold for that case but were omitted due to space constraints. Instead, we use simulation results to explore this case. In particular, to evaluate the relative benefits of local generation and workload shifting in practice, we can contrast Figures 4–7. These simulation results highlight that local generation is crucial, in order to provide responses to warning signals from the utility; but at the same time, even when local generation is present, workload shifting can provide significant cost savings, and can lead to a significant reduction in the amount of local generation needed (and thus emissions).

More specifically, compared with the case of no local generation, the use of local generation can help reduce the coincident peak costs; however one must be careful when using local generation to correct for prediction error since this added cost is not worth it unless the prediction error is extreme. The aggregate effect is perhaps smaller than expected, and can be seen by comparing Figure 5(e) with 7(e) and Figure 6(e) with 4(k). As discussed in Section 4, the benefit of local generation depends on the number of warnings, the local generation cost, and the prediction error. With fewer warnings and cheaper local generation, local generators can help reduce costs more. However, this benefit comes with higher emissions (5-10% in the experiments) since local generators are usually not environmentally friendly. This can be seen from the emission comparison between Figures 5(f) and 7(f), and Figures 6(f) and 4(l). Importantly, renewable generation can help reduce both energy costs and emissions significantly, especially when combined with workload management. This can be seen from cost and emission comparisons across Figures 5 and 6, and Figures 7 and 4.

5.2.3. Sensitivity to prediction errors

The final issue that we seek to understand using our experiments is the impact of prediction errors. We have already provided an analytic characterization of the impact of prediction errors on workload and renewable generation



(a) Prediction: one week plan (b) Prediction: one day plan (c) Robust: one week plan

(d) Robust: one day plan (e) Energy costs (f) Emissions

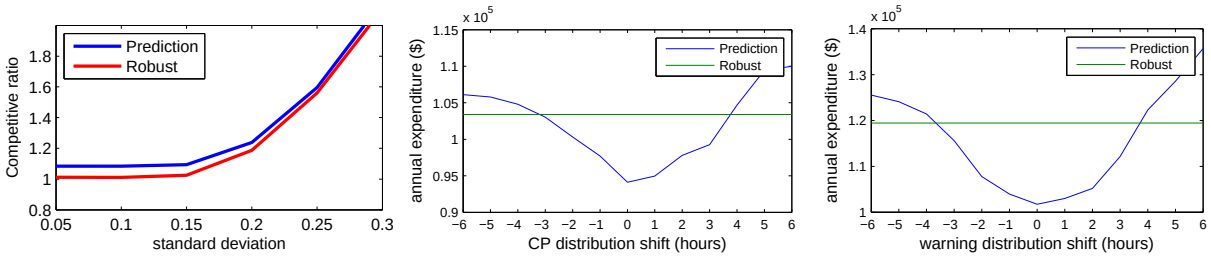
Figure 7. Comparison of energy costs and emissions for a data center with a local diesel generator, but without local PV generation. (a)-(d) show the plans computed by our algorithms.

in Section 4 and so (due to limited space) we only briefly comment on numerical results corroborating our analysis here – Figure 8(a) shows the growth of the competitive ratio as a function of the standard deviation of the prediction error. Recall that all results in Figures 4–7 incorporate prediction errors as well.

More importantly, we focus this section on coincident peak and warning prediction errors. Figure 8 studies this issue. In this figure, the predictions used by *Prediction* are manipulated to create inaccuracies. In particular, the predictions calculated via the historical data are shifted earlier/later by up to 6 hours, and the corresponding energy costs and emissions are shown. Of course, the costs and emissions of *Robust* are unaffected by the change in the predictions; however the costs and emissions of *Prediction* change dramatically. In particular, *Prediction* becomes worse than *Robust* if the shift (and the error) in the prediction distribution is larger than 3.5 hours.

6. Concluding Remarks

Our goal in this paper is to provide algorithms to plan for workload shifting and local generation usage at a data center participating in a CPP demand response program with uncertainties in coincident peak and warnings, workload demand and renewable generation. To this end, we have obtained and characterized a 26-year data set from the CPP program run by Fort Collins Utilities, Colorado. This characterization provides important new insights about CPP programs that can be useful for data center demand response algorithms. Using these insights, we have presented two approaches for designing algorithms for workload management and local generation planning at a data center participating in a CPP program. In particular, we have presented a stochastic optimization based algorithm that seeks to minimize the expected energy expenditure using predictions about when the coincident peak and corresponding warnings will occur, workload demand and renewable generation, and another robust optimization based algorithm designed to provide minimal worst case guarantees on energy expenditure given all uncertainties. Finally, we have



(a) Workload/Renewable error (b) Spread out CPs (c) Spread out warnings

Figure 8. Sensitivity analysis of "Prediction" and "Robust" algorithms with respect to (a) workload and renewable generation prediction error and (b) & (c) coincident peak and warning prediction errors. In all cases, the data center considered has a local diesel generator, but no local PV installation.

evaluated these algorithms using detailed, real world trace-based numerical simulation experiments. These experiments highlight that the use of both workload shifting and local generation are crucial in order for a data center to minimize its energy costs and emissions.

There are a number of future research directions that build on the work in this paper. In particular, an interesting direction is to adapt the algorithms presented here in order to incorporate energy storage at the data center. More generally, Internet-scale systems are typically provided by a geographically distributed data centers, and so it would be interesting to understand how the "geographical load balancing" performed by such systems interacts with coincident peak pricing. This "moving bits, not watts" scheme can significantly reduce local power network pressure without adding further load to the (possibly already) congested transmission network. Additionally, CPP programs are just one example of demand response programs. Though CPP programs are currently the most common form of demand response program, a number of new programs are emerging. It is important to understand how each of these programs, e.g., [49], interact with data center planning.

References

[1] National Institute of Standards and Technology, "NIST framework and roadmap for smart grid interoperability standards." NIST Special Publication 1108, 2010.
[2] Department of Energy, "The smart grid: An introduction."
[3] Federal Energy Regulatory Commission, "National assessment of demand response potential." 2009.
[4] NY Times, "Power, Pollution and the Internet."
[5] G. Ghatikar, V. Ganti, N. Matson, and M. Piette, "Demand response opportunities and enabling technologies for data centers: Findings from field studies," 2012.
[6] "Report to congress on server and data center energy efficiency." 2007.
[7] J. Koomey, "Growth in data center electricity use 2005 to 2010," *Oakland, CA: Analytics Press. August*, vol. 1, p. 2010, 2011.
[8] www.fcgov.com/utilities/business/rates/electric/coincident-peak.
[9] http://www.marketwire.com/press-release/webair-enernoc-turn-data-centers-into-virtual-power-plants-through-demand-response-1408389.htm.
[10] A. Gandhi, Y. Chen, D. Gmach, M. Arlitt, and M. Marwah, "Minimizing data center sla violations and power consumption via hybrid resource provisioning," in *Proc. of IGCC*, 2011.
[11] Y. Chen, D. Gmach, C. Hyser, Z. Wang, C. Bash, C. Hoover, and S. Singhal, "Integrated management of application performance, power and cooling in data centers," in *Proc. of NOMS*, 2010.
[12] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," in *Proc. of INFOCOM*, 2011.
[13] S. Govindan, J. Choi, B. Urgaonkar, A. Sivasubramaniam, and A. Baldini, "Statistical profiling-based techniques for effective power provisioning in data centers," in *Proc. of EuroSys*, 2009.
[14] J. Choi, S. Govindan, B. Urgaonkar, and A. Sivasubramaniam, "Profiling, prediction, and capping of power consumption in consolidated environments," in *MASCOTS*, 2008.

[15] J. Heo, P. Jayachandran, I. Shin, D. Wang, T. Abdelzaher, and X. Liu, "Optituner: On performance composition and server farm energy minimization application," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 11, pp. 1871–1878, 2011.
[16] A. Verma, G. Dasgupta, T. Nayak, P. De, and R. Kothari, "Server workload analysis for power minimization using consolidation," in *USENIX ATC*, 2009.
[17] D. Meisner, C. Sadler, L. Barroso, W. Weber, and T. Wenisch, "Power management of online data-intensive services," in *Proc. of ISCA*, 2011.
[18] Q. Zhang, M. Zhani, Q. Zhu, S. Zhang, R. Boutaba, and J. Hellerstein, "Dynamic energy-aware capacity provisioning for cloud computing environments," in *ICAC*, 2012.
[19] H. Xu and B. Li, "Cost efficient datacenter selection for cloud services," 2012.
[20] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely, "Data centers power reduction: A two time scale approach for delay tolerant workloads," in *Proc. of INFOCOM*, pp. 1431–1439, 2012.
[21] R. Urgaonkar, B. Urgaonkar, M. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *Proc. of the ACM Sigmetrics*, 2011.
[22] D. Irwin, N. Sharma, and P. Shenoy, "Towards continuous policy-driven demand response in data centers," *Computer Communication Review*, vol. 41, no. 4, 2011.
[23] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *Proc. of ACM Sigmetrics*, 2012.
[24] K. Le, O. Bilgir, R. Bianchini, M. Martonosi, and T. D. Nguyen, "Capping the brown energy consumption of internet services at low cost," in *Proc. IGCC*, 2010.
[25] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Greening geographical load balancing," in *Proc. ACM Sigmetrics*, 2011.
[26] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment," in *Proc. of INFOCOM*, 2010.
[27] P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford, "Donar: decentralized server selection for cloud services," in *Proc. of ACM Sigcomm*, 2010.
[28] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Geographical load balancing with renewables," in *Proc. ACM GreenMetrics*, 2011.
[29] M. Lin, Z. Liu, A. Wierman, and L. Andrew, "Online algorithms for geographical load balancing," in *Proc. of IGCC*, 2012.
[30] D. Meisner, J. Wu, and T. Wenisch, "Bighouse: A simulation infrastructure for data center systems," in *Proc. of ISPASS*, pp. 35–45, 2012.
[31] L. Barroso and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis Lectures on Computer Architecture*, vol. 4, no. 1, pp. 1–108, 2009.
[32] www.ge-energy.com.
[33] http://www.apple.com/environment/renewable-energy.
[34] B. Urgaonkar, G. Pacifici, P. Shenoy, M. Spreitzer, and A. Tantawi, "An analytical model for multi-tier internet services and its applications," in *Proc. of ACM Sigmetrics*, 2005.
[35] Y. Chen, S. Alspaugh, and R. Katz, "Interactive analytical processing in big data systems: a cross-industry study of mapreduce workloads," *Proc. of VLDB*, 2012.
[36] M. Zaharia, D. Borthakur, J. Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, "Job scheduling for multi-user mapreduce clusters," *UCB/EECS-2009-55*, 2009.
[37] T. Breen, E. Walsh, J. Punch, C. Bash, and A. Shah, "From chip to cooling tower data center modeling: Influence of server inlet temperature and temperature rise across cabinet," *Journal of Electronic Packaging*, vol. 133, no. 1, 2011.
[38] C. Patel, R. Sharma, C. Bash, and A. Beitelmal, "Energy flow in the information technology stack," in *Proc. of IMECE*, 2006.
[39] EPA, "US Emission Standards for Nonroad Diesel Engines." www.dieselnet.com/standards/us/nonroad.php.
[40] C. Ren, D. Wang, B. Urgaonkar, and A. Sivasubramaniam, "Carbon-aware energy capacity planning for datacenters," in *MASCOTS*, pp. 391–400, IEEE, 2012.
[41] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *Proc. of ACM Sigcomm*, 2009.
[42] C. Stewart and K. Shen, "Some joules are more precious than others: Managing renewable energy in the datacenter," in *Proc. of HotPower*, 2009.
[43] K. Le, R. Bianchini, M. Martonosi, and T. Nguyen, "Cost-and energy-aware load distribution across data centers," *Proceedings of HotPower*, 2009.
[44] Í. Goiri, K. Le, T. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenhadoop: leveraging green energy in data-processing frameworks," in *Proc. of EuroSys*, 2012.
[45] N. Deng, C. Stewart, J. Kelley, D. Gmach, and M. Arlitt, "Adaptive green hosting," in *Proceedings of ICAC*, 2012.
[46] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *Proc. of SmartGridComm*, 2011.
[47] Y. Becerra, D. Carrera, and E. Ayguade, "Batch job profiling and adaptive profile enforcement for virtualized environments," in *Proc. of ICPDNP*, 2009.
[48] J. Choi, S. Govindan, B. Urgaonkar, and A. Sivasubramaniam, "Power consumption prediction and power-aware packing in consolidated environments," *IEEE Transactions on Computers*, vol. 59, no. 12, 2010.
[49] D. Aikema, R. Simmonds, and H. Zareipour, "Data centres in the ancillary services market,"

Appendix A. Proofs

In this appendix we include proofs for bounds on the competitive ratio of our both algorithms in Section 4. Because the proof of Theorem 1 uses simplified versions of many parts of the proof of Theorem 2, we start with the proof of Theorem 2 and then describe how to specialize the approach to Theorem 1.

To prove Theorem 2, we start with some notation and simple observations. First, in this context, the offline optimal is defined as follows: $(\mathbf{b}^*, \mathbf{g}^*) \in \mathbf{argmin}_{\mathbf{b},\mathbf{g}} f^*(\mathbf{e}, \mathbf{g})$, where $f^*(\mathbf{e}, \mathbf{g}) \equiv \Sigma_t p(t)e(t) + p_p \mathbf{max}_t e(t) + p_{cp} e(t_{cp}) + p_g \Sigma_t g(t)$. Here $\mathbf{b}$ stands for the workload management, and $\mathbf{g}$ denotes the local backup generator usage, $e(t) = (d(t) - r(t) - g(t))^+$ is the grid power usage, we assume the offline optimal have perfect knowledge of $d(t)$, $r(t)$, and when coincidental peak occurs.

In contrast, the plan derived from Algorithm 2, denoted by $(\hat{\mathbf{e}}_1^w, \mathbf{g}_1^w)$, minimizes

$$f^w(\hat{\mathbf{e}}, \mathbf{g}) \equiv \Sigma_t p(t)\hat{e}(t) + \left(p_p + \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\right)\mathbf{max}_t \hat{e}(t) + p_g \Sigma_t g(t)$$

using prediction of workload $\hat{d}(t)$ and prediction of renewable generation $\hat{r}(t)$ without any knowledge of coincidental peak (CP) or warnings except $\bar{W}$. Here $\hat{e}(t) = (\hat{d}(t) - \hat{r}(t) - g(t))^+$. In addition, Algorithm 2 uses minimal local generation to remove harmful prediction error when (4) occurs, i.e., $g_\varepsilon^w(t) = \max\{0, \min\{e^w(t), \varepsilon_d \hat{d}^w(t) - \varepsilon_r \hat{r}(t)\}\}$. Also, Algorithm 2 uses local generation whenever warnings are received, i.e., $g_2^w(t) = I_{\{t \in W\}} e_1^w(t), \forall t$, where $I_{\{t \in W\}}$ is the indicator function, which equals to 1 if t is a time when warning is received and 0 otherwise and $e_1^w(t) = (d_1^w(t) - r(t) - g_1^w(t) - g_\varepsilon^w(t))^+$. Therefore the real grid power usage at time t is $e^w(t) \le \hat{e}_1^w(t) - g_2^w(t)$, and local power generation is $g^w(t) = g_1^w(t) + g_\varepsilon^w(t) + g_2^w(t), \forall t$. Note here $(\hat{\mathbf{e}}_1^w, \mathbf{g}_1^w)$ is the day-ahead plan, while $(\mathbf{e}^w, \mathbf{g}^w)$ is the real grid power consumption and local generation after using local generation to compensate for underestimation and during warning periods.

Proof of Theorem 2. Note that f^* and f^w are optimizations using different data (f^* uses perfect knowledge of $d(t)$ and $r(t)$, while f^w uses prediction $\hat{d}(t)$ and $\hat{r}(t)$), to bridge this gap, we first observe the following:

$$f^*(\mathbf{e}^*, \mathbf{g}^*) \ge f^*(\hat{\mathbf{e}}^*, \mathbf{g}^* + \mathbf{g}_\varepsilon^*) - p_g \Sigma_t g_\varepsilon^*(t) \tag{A.1}$$

where $\hat{\mathbf{e}}^*$ is the optimizer of f^* using prediction $\hat{d}(t)$ and $\hat{r}(t)$, and $\mathbf{g}_\varepsilon^*$ is defined in a similar way to $\mathbf{g}_\varepsilon^w$, $g_\varepsilon^*(t) = \max\{0, \min\{\hat{e}^*(t), \varepsilon_d \hat{d}(t) - \varepsilon_r \hat{r}(t)\}\}$ which removes all the harmful prediction errors. The right hand side of the inequality is essentially evaluating the same objective using prediction, but is given $\mathbf{g}_\varepsilon^*$ of local power for free. As $\mathbf{g}_\varepsilon^*$ removes all harmful effects of prediction, using prediction will not increase the objective.

The key step is to bound $\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^w(\hat{\mathbf{e}}_1^w, \mathbf{g}_1^w)]$ in terms of $\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\hat{\mathbf{e}}^*, \mathbf{g}^* + \mathbf{g}_\varepsilon^*)]$

$$
\begin{aligned}
&\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\hat{\mathbf{e}}^*, \mathbf{g}^* + \mathbf{g}_\varepsilon^*)] \\
&= \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^w(\hat{\mathbf{e}}^*, \mathbf{g}^* + \mathbf{g}_\varepsilon^*))] - \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t \hat{e}^*(t)] + p_{cp}\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\hat{e}^*(t_{cp})] \\
&\geq \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^w(\hat{\mathbf{e}}^*, \mathbf{g}^* + \mathbf{g}_\varepsilon^*))] - \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t \hat{e}^*(t)] \\
&\geq \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^w(\hat{\mathbf{e}}_1^w, \mathbf{g}_1^w)] - \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t \hat{e}^*(t)] \\
&\geq \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^w(\hat{\mathbf{e}}_1^w, \mathbf{g}_1^w + \mathbf{g}_\varepsilon^w)] - p_g\Sigma_t\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[g_\varepsilon^w(t)] - \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t \hat{e}^*(t)] \\
&\geq \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\mathbf{e}^w, \mathbf{g}^w)] - p_g\Sigma_t\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[g_\varepsilon^w(t)] - \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t \hat{e}^*(t)] \qquad \text{(A.2)}
\end{aligned}
$$

Here the first inequality holds because $p_{cp}\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\hat{e}^*(t_{cp})] \geq 0$. The second inequality is from the optimality of $(\hat{\mathbf{e}}_1^w, \mathbf{g}_1^w)$ in minimizing $f^w(\mathbf{e}, \mathbf{g})$. However, the last inequality is more involved.

We show the last step of (A.2) by first writing out the day-ahead plan $\hat{e}_1^w(t) = \left(\hat{d}_1^w(t) - \hat{r}(t) - g_1^w(t)\right)^+$, and the actual power demand $e^w(t) = \left(d_1^w(t) - r(t) - g_1^w(t) - g_\varepsilon^w(t) - g_2^w(t)\right)^+$. Furthermore, denote $e_2^w(t)$ as the electricity demand of Algorithm 2 without using local generation to respond to CP warning. Then $e^w(t) = e_2^w(t) - g_2^w(t)$, and $g_2^w(t) = e_2^w(t)I_{\{t\in W\}}$, so we have

$$
e_2^w(t) = \left(d_1^w - r(t) - g_1^w(t) - g_\varepsilon^w(t)\right)^+ \leq \left(\hat{d}_1^w(t) - \hat{r}(t) - g_1^w(t)\right)^+ = \hat{e}_1^w(t)
$$

Hence $e^w(t) = e_2^w(t) - g_2^w(t) \leq \hat{e}_1^w(t) - g_2^w(t)$. Next, we bound $f^*(\mathbf{e}^w, \mathbf{g}^w)$ as follows.

$$
\begin{aligned}
f^*(\mathbf{e}^w, \mathbf{g}^w) &= f^*(\mathbf{e}^w, \mathbf{g}_1^w + \mathbf{g}_\varepsilon^w + \mathbf{g}_2^w) \\
&= \Sigma_t p(t)e^w(t) + p_p\mathbf{max}_t e^w(t) + p_{cp}e^w(t_{cp}) + p_g\Sigma_t g^w(t) \\
&= \Sigma_{t\notin W} p(t)e_2^w(t) + p_p\mathbf{max}_{t\notin W} e_2^w(t) + p_g\left(\Sigma_t(g_1^w(t) + g_\varepsilon^w(t)) + \Sigma_{t\in W} e_2^w(t)\right) \\
&\leq \Sigma_t p(t)\hat{e}_1^w(t) + p_p\mathbf{max}_t \hat{e}_1^w(t) + p_g\Sigma_t(g_1^w(t) + g_\varepsilon^w(t)) + \Sigma_{t\in W}(p_g - p(t))\hat{e}_1^w(t) \\
&\leq \Sigma_t p(t)\hat{e}_1^w(t) + p_p\mathbf{max}_t \hat{e}_1^w(t) + p_g\Sigma_t(g_1^w(t) + g_\varepsilon^w(t)) + \bar{W}(p_g - \mathbf{min}_t p(t))\mathbf{max}_t \hat{e}_1^w(t) \\
&= f^w(\hat{\mathbf{e}}_1^w, \mathbf{g}_1^w + \mathbf{g}_\varepsilon^w) \qquad \text{(A.3)}
\end{aligned}
$$

The second equality is because $g_2^w(t) = I_{\{t\in W\}}e_2^w(t), \forall t$. The first inequality is from $\mathbf{max}_{t\notin W} e_2^w(t) \leq \mathbf{max}_t e_2^w(t)$ and $e_2^w(t) \leq \hat{e}_1^w(t)$. The second inequality holds because $\Sigma_{t\in W}(p_g - p(t))\hat{e}_1^w(t) \leq \Sigma_{t\in W}(p_g - \mathbf{min}_t p(t))\hat{e}_1^w(t)$
$= \left(p_g - \mathbf{min}_t p(t)\right)\Sigma_{t\in W}\hat{e}_1^w(t) \leq \left(p_g - \mathbf{min}_t p(t)\right)\Sigma_{t\in W}\mathbf{max}_t \hat{e}_1^w(t) \leq \bar{W}(p_g - \mathbf{min}_t p(t))\mathbf{max}_t \hat{e}_1^w(t)$.

Finally, we can combine (A.1) and (A.2) to obtain

$$
\begin{aligned}
&\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\mathbf{e}^*, \mathbf{g}^*)] \\
&\geq \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\hat{\mathbf{e}}^*, \mathbf{g}^* + \mathbf{g}_\varepsilon^*)] - p_g\Sigma_t\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[g_\varepsilon^*(t)] \\
&\geq \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r} f^*(\mathbf{e}^w, \mathbf{g}^w) - p_g\Sigma_t\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[g_\varepsilon^w(t) + g_\varepsilon^*(t)] - \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t \hat{e}^*(t)] \\
&\geq \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\mathbf{e}^w, \mathbf{g}^w)] - p_g\sigma\Sigma_t\left(\frac{\hat{d}^w(t) + \hat{d}^*(t)}{2} + \hat{r}(t)\right) - \bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}\mathbf{max}_t \hat{e}^*(t), \qquad \text{(A.4)}
\end{aligned}
$$

where (A.4) derives from the following

$$
\begin{aligned}
&\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[g^w_\varepsilon(t) + g^*_\varepsilon(t)] \\
&= \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\max\{0, \min\{e^w(t), \varepsilon_d\hat{d}^w(t) - \varepsilon_r\hat{r}(t)\}\} + \max\{0, \min\{e^*(t), \varepsilon_d\hat{d}^*(t) - \varepsilon_r\hat{r}(t)\}\}] \\
&\le \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[(\varepsilon_d\hat{d}^w(t) - \varepsilon_r\hat{r}(t))^+] + \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[(\varepsilon_d\hat{d}^*(t) - \varepsilon_r\hat{r}(t))^+] \\
&= \mathbb{E}[\varepsilon^w(t)^+] + \mathbb{E}[\varepsilon^*(t)^+] \qquad \left(\text{let } \varepsilon^w(t) = \varepsilon_d\hat{d}^w(t) - \varepsilon_r\hat{r}(t), \varepsilon^*(t) = \varepsilon_d\hat{d}^*(t) - \varepsilon_r\hat{r}(t)\right) \\
&\le \frac{1}{2}\sigma_{\varepsilon^w(t)} + \frac{1}{2}\sigma_{\varepsilon^*(t)} \\
&= \frac{1}{2}\left(\sqrt{\hat{d}^w(t)^2\sigma_d^2 + \hat{r}(t)^2\sigma_r^2} + \sqrt{\hat{d}^*(t)^2\sigma_d^2 + \hat{r}(t)^2\sigma_r^2}\right) \\
&\le \frac{1}{2}\left((\hat{d}^w(t) + \hat{r}(t))\max(\sigma_d, \sigma_r) + (\hat{d}^*(t) + \hat{r}(t))\max(\sigma_d, \sigma_r)\right) \\
&= \left(\frac{\hat{d}^w(t) + \hat{d}^*(t)}{2} + \hat{r}(t)\right)\sigma
\end{aligned} \tag{A.5}
$$

The second last equality holds because ε_d and ε_r are independent, and the last inequality holds because $\hat{d}(t)$ and $\hat{r}(t)$ are nonnegative.

The key is the second inequality, as the cases for $\varepsilon^w(t)$ and $\varepsilon^*(t)$ are the same, we just need to show this inequality holds for any $\varepsilon(t)$ has zero mean and fixed variance $\sigma^2_{\varepsilon(t)}$. Note that $\varepsilon(t) = \varepsilon(t)^+ - \varepsilon(t)^-$, hence $\mathbb{E}[\varepsilon(t)] = 0 \Rightarrow \mathbb{E}[\varepsilon(t)^+] = \mathbb{E}[\varepsilon(t)^-]$. It follows that

$$
\begin{aligned}
\sigma^2_{\varepsilon(t)} &= \mathbb{E}[\varepsilon(t)^2] \\
&= \mathbb{E}[(\varepsilon(t)^+)^2] + \mathbb{E}[(\varepsilon(t)^-)^2] - 2\mathbb{E}[\varepsilon(t)^+\varepsilon(t)^-] \\
&= \mathbb{E}[(\varepsilon(t)^+)^2] + \mathbb{E}[(\varepsilon(t)^-)^2] \\
&\ge \frac{\mathbb{E}[\varepsilon(t)^+]^2}{\mathbb{P}(\varepsilon(t) \ge 0)} + \frac{\mathbb{E}[\varepsilon(t)^-]^2}{\mathbb{P}(\varepsilon(t) < 0)} \\
&= \mathbb{E}[\varepsilon(t)^+]^2\left(\frac{1}{\mathbb{P}(\varepsilon(t) \ge 0)} + \frac{1}{1 - \mathbb{P}(\varepsilon(t) \ge 0)}\right) \\
&= \mathbb{E}[\varepsilon(t)^+]^2\left(\frac{1}{(\mathbb{P}(\varepsilon(t) \ge 0))(1 - \mathbb{P}(\varepsilon(t) \le 0))}\right) \\
&\ge 4\mathbb{E}[\varepsilon(t)^+]^2
\end{aligned}
$$

Rearranging, we have $\mathbb{E}[\varepsilon(t)^+] \le \frac{1}{2}\sigma_{\varepsilon(t)}$. The last inequality attains equality when $\mathbb{P}(\varepsilon(t)^+ \ge 0) = \mathbb{P}(\varepsilon(t)^- < 0) = 1/2$. The third equality follows because $\varepsilon(t)^+$ and $\varepsilon(t)^-$ cannot be simultaneously non-zero. The first inequality

follows because

$$\begin{aligned}
&\mathbb{E}[(\varepsilon(t)^+)^2]\mathbb{P}(\varepsilon(t) \geq 0) \\
&= \int_0^\infty x^2 dF_{\varepsilon(t)}(x) \int_0^\infty 1 dF_{\varepsilon(t)}(x) \\
&\geq \left(\int_0^\infty x \cdot 1 dF_{\varepsilon(t)}(x) \right)^2 \\
&= \mathbb{E}[\varepsilon(t)^+]^2 \\
&\Rightarrow \mathbb{E}[(\varepsilon(t)^+)^2] \geq \frac{\mathbb{E}[\varepsilon(t)^+]^2}{\mathbb{P}(\varepsilon(t) \geq 0)}
\end{aligned}$$

The first inequality follows from Cauchy-Schwarz inequality, and the inequality attains equality when the distribution of $\varepsilon(t)^+$ is a point mass. By similar argument we can show that $\mathbb{E}[\varepsilon(t)^-]^2 \geq \frac{\mathbb{E}[\varepsilon(t)^-]^2}{\mathbb{P}(\varepsilon(t)<0)}$, and equality is attained when the distribution of $\varepsilon(t)^-$ is a point mass.

Using the observation above and the previous observation that $\mathbb{P}(\varepsilon(t)^+ \geq 0) = \mathbb{P}(\varepsilon(t)^- < 0) = 1/2$, we can see that $\mathbb{E}[\varepsilon(t)^+] = \frac{1}{2}\sigma_{\varepsilon(t)}$ when the distribution of $\varepsilon(t)$ is two equal point masses located at $\sigma_{\varepsilon(t)}$ and $\sigma_{\varepsilon(t)}$ respectively.

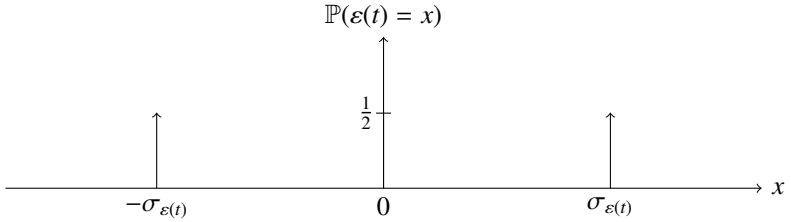


Figure A.9. Illustration of pdf of $\varepsilon(t)$ that attains $\mathbb{E}[\varepsilon(t)^+] = \frac{1}{2}\sigma_{\varepsilon(t)}$ for $\mathbb{E}[\varepsilon(t)] = 0$ and $Var(\varepsilon(t)) = \sigma_{\varepsilon(t)}$.

Finally, combining the above, we can compute the competitive ratio as follows

$$\begin{aligned}
&\frac{\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\mathbf{e}^w, \mathbf{g}^w)]}{\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\mathbf{e}^*, \mathbf{g}^*)]} \\
&\leq 1 + \frac{\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t e^*(t)] + p_g \sigma \Sigma_t \left(\frac{\hat{d}^w(t)+\hat{d}^*(t)}{2} + \hat{r}(t)\right)}{\Sigma_t p(t)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[e^*(t)] + p_p \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t e^*(t)] + p_{cp}\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[e^*(t_{cp})] + p_g \Sigma_t g^*(t)} \\
&\leq 1 + \frac{\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)}{\Sigma_t p(t)\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[e^*(t)]/\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t e^*(t)] + p_p} + B\sigma, \qquad \left(B = \frac{p_g \Sigma_t (\frac{\hat{d}^w(t)+\hat{d}^*(t)}{2} + \hat{r}(t))}{\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[f^*(\mathbf{e}^*, \mathbf{g}^*)]}\right) \qquad \text{(A.6)} \\
&\leq 1 + \frac{\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)}{\mathbf{min}_t p(t)\Sigma_t \mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[e^*(t)]/\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}[\mathbf{max}_t e^*(t)] + p_p} + B\sigma \\
&= 1 + \frac{\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)}{T\mathbf{min}_t p(t)/PMR^* + p_p} + B\sigma \\
&\leq 1 + \frac{\bar{W}\left(p_g - \mathbf{min}_t p(t)\right)}{p_p} + B\sigma
\end{aligned}$$

It remains to show that no online algorithm can have competitive ratio smaller than $(1 + \frac{\bar{W}(p_g - \min_t p(t))}{p_p})$ even with perfect information of workload and renewable generation. To prove this, we use the instance summarized in Figure A.10.

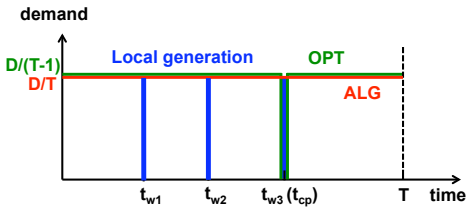


Figure A.10. Instance for lower bounding the competitive ratio for setting with local generation.

In this instance, PUE is the same across all time slots and small. There is no local renewable supply or interactive workload. The total flexible workload demand is D. The (discrete) time horizon is [1,T], where $t_{wi}, i = 1, ..., W$ are the time slots with warnings (three warnings are shown in the figure) and the total number of warnings is W with bound $\bar{W} \geq W$ known to the online algorithm. The final coincident peak hour is t_{cp} and it is among the warnings (t_{w3} in the figure). The usage-based electricity price $p(t) = p, \forall t$ and is much smaller than p_p and p_{cp}. Also, in this instance, $\frac{p_p}{T-1} \leq p_g$ (using local generation is more expensive than demand shifting and paying (slightly) increased peak demand charging) and $p_g \leq p_{cp}$, which are common in practice.

In this setting, the offline optimal solution plans according to the green curve: it does not use the coincident peak time slot but spreads the demand evenly across the other $T - 1$ time slots. The cost of the offline optimal solution is therefore $f^*(\mathbf{e}^*, \mathbf{g}^*) = pD + p_p \frac{D}{T-1}$.

In contrast, any online algorithm can at best plan according to the red curve: spreading the workload evenly among all T time slots and using local generation when warnings are received. To see this, note that there is no benefit to spreading the workload unevenly since that increases local generation usage for the worst-case instance and possibly the peak charging, while not saving any usage based cost. The cost of the best online non-adaptive solution is therefore $f^*(\mathbf{e}^{ALG}, \mathbf{g}^{ALG}) = pD + p_p \frac{D}{T} + W\left(p_g - p\right)\frac{D}{T}$. The best competitive ratio is therefore:

$$\begin{aligned}
\frac{f^*(\mathbf{e}^{ALG}, \mathbf{g}^{ALG})}{f^*(\mathbf{e}^*, \mathbf{g}^*)} &= \frac{pD + p_p \frac{D}{T} + W\left(p_g - p\right)\frac{D}{T}}{pD + p_p \frac{D}{T-1}} \\
&= 1 + \frac{-p_p \frac{D}{T(T-1)} + W\left(p_g - p\right)\frac{D}{T}}{pD + p_p \frac{D}{T-1}} \\
&= 1 + \frac{W(p_g - p) - \frac{p_p}{T-1}}{pT + p_p \frac{T}{T-1}}
\end{aligned}$$

As $T \to \infty$, taking the usage cost pT as the same or smaller order of magnitude as the peak cost p_p, this becomes

$$1 + \frac{W(p_g - p)}{pT + p_p}$$

The above matches the bound in equation (A.6) when $W = \bar{W}$, which completes the proof. □

Proof Sketch of Theorem 1. The proof of Theorem 1 is similar in structure to that of Theorem 2, only simpler. Thus, we outline only the main steps and highlight the similarities with the proof of Theorem 2. In particular, the following provides the major steps needed to bridge the expected cost of Algorithm 1 and the cost of the offline algorithm with exact IT demand and renewable generation knowledge:

$$\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r,\hat{W}}\left[f(\mathbf{e}^*,\mathbf{g}^*)\right]$$

$$\geq \mathbb{E}_{\hat{W}}\left[\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}\left[f(\hat{\mathbf{e}}^*,\mathbf{g}^*+\mathbf{g}^*_\epsilon)-p_g\sum_{t=1}^{T}g^*_\varepsilon(t)\right]\right] \tag{A.7a}$$

$$= \mathbb{E}_{\hat{W}}\left[\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}\left[f^s(\hat{\mathbf{e}}^*,\mathbf{g}^*+\mathbf{g}^*_\epsilon)\right]-p_g\sum_{t=1}^{T}\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}\left[g^*_\varepsilon(t)\right]\right] \tag{A.7b}$$

$$\geq \mathbb{E}_{\hat{W}}\left[\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}\left[f^s(\mathbf{e}^s,\mathbf{g}^s_1)\right]-\frac{1}{2}\sigma p_g\sum_{t=1}^{T}\left(\hat{d}^*(t)+\hat{r}(t)\right)\right] \tag{A.7c}$$

$$\geq \mathbb{E}_{\hat{W}}\left[\mathbb{E}_{\hat{\xi}_d,\hat{\xi}_r}\left[f(\mathbf{e}^s,\mathbf{g}^s)\right]-\frac{1}{2}\sigma p_g\sum_{t=1}^{T}\left(\hat{d}^*(t)+\hat{r}(t)\right)-\frac{1}{2}\sigma p_g\sum_{t=1}^{T}\left(\hat{d}^s(t)+\hat{r}(t)\right)\right] \tag{A.7d}$$

It is easy to see that the theorem follows from this general approach, but of course each step requires some effort to justify. However, the justification of each step parallels calculations from the proof of Theorem 2. In particular, (A.7a) is parallel to (A.1), (A.7b) is because $f(\cdot)$ and $f^s(\cdot)$ are equivalent when taking expectation, (A.7c) is parallel to (A.5), and (A.7d) is parallel to (A.2). Since the verification of these is simpler than in the case of Theorem 2, we omit the details. □