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

Zhenhua Liu
California Institute of Technology
Pasadena, CA, USA
zliu2@caltech.edu

Adam Wierman
California Institute of Technology
Pasadena, CA, USA
adamw@caltech.edu

Yuan Chen
HP Labs
Palo Alto, CA, USA
yuan.chen@hp.com

Benjamin Razon
California Institute of Technology
Pasadena, CA, USA
ben@caltech.edu

Niangjun Chen
California Institute of Technology
Pasadena, CA, USA
ncchen@caltech.edu

## 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 the high and increasing energy consumption and the flexibility in demand management in data centers compared to conventional industrial facilities. In this extended abstract we briefly describe recent work in [1] on two demand response schemes to reduce a data center's peak loads and energy expenditure: workload shifting and the use of local power generations. In [1], 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. 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% in the Fort Collins Utilities' case) compared to either alone.

## Categories and Subject Descriptors

C.0 [**Computer Systems Organization**]: General

## Keywords

Demand response, coincident peak pricing, data center, workload shifting, online algorithm

## 1. INTRODUCTION

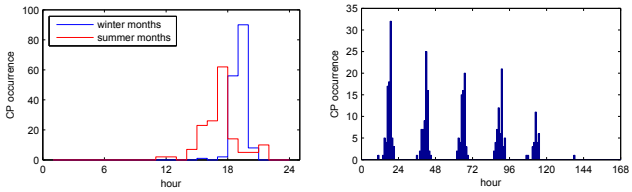
Demand response (DR) programs seek to provide incentives to induce dynamic demand management of customers' electricity load in response to power supply conditions, for example, reducing their power consumption in response to a peak load warning signal or request from the utility. The National Institute 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 [2, 3]. 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 [4]. 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 [5, 6, 7]. Second, data centers are highly automated and monitored, and so there is the potential for a high-degree of responsiveness. Third, many workloads in data centers are delay tolerant, which enables significant flexibility for optimizing power demand. Finally, local power generation, e.g., traditional backup generators and newer renewable power installations, can help shape the power demand from the grid. 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 in data centers, the current reality is that industry data centers seemingly perform little, if any, demand response [5, 6]. One popular demand response programs available is Coincident Peak Pricing (CPP), which is required for medium and large industrial consumers in many regions. These programs work by charging a very high price for usage during the coincident peak hour, which is the hour when the most electricity is requested from the utility's wholesale electric supplier. This rate is often over 200 times higher than the base rate, so it is common for the coincident peak charges to account for over 23% of a customer's electric bill according to Fort Collins Utilities. 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.

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. 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


*SIGMETRICS'13,* June 17-21, 2013, Pittsburgh, PA, USA.
ACM 978-1-4503-1900-3/13/06.




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

**Figure 1: Occurrence of coincident peak. (a) Empirical frequency of occurrences on the time of day (b) Empirical frequency of occurrences over the week.**

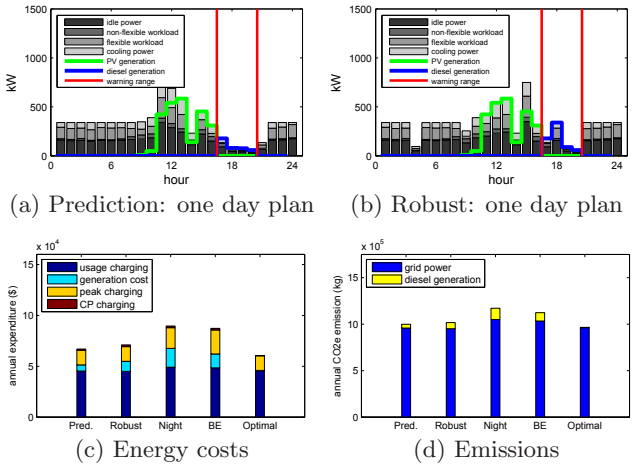
not meeting Environmental Protection Agency (EPA) emissions standards [5]. So, from an environmental perspective this form of response is far from ideal. Further, running a backup generator can be expensive. Alternatively, providing demand response via shifting workload can be more cost effective. A challenge with workload shifting is that we need to ensure that the Service Level Agreements (SLAs), e.g., completion deadlines, remain satisfied.

## 2. OVERVIEW OF RESULTS

In this abstract, we briefly discuss the main contributions of the work in [1]. First, we present **a detailed characterization study of coincident peak pricing** and provide insight about its properties. We characterize 26 years' coincident peak pricing data from Fort Collins Utilities. The data highlights a number of important observations. For example, the data set shows that the coincident peak occurrence has a strong diurnal pattern that differs considerably during different days of the week and across seasons, as shown in Figure 1. Further, the data highlights that coincident peak warnings are highly reliable – only twice did the coincident peak not occur during a warning hour.

Second, we develop **two online algorithms for avoiding the coincident peak and reducing the energy expenditure using workload shifting and local power generation.** The uncertainty of the occurrence of the coincident peak hour presents significant challenges for workload scheduling and local generation planning. For example, traditional workload scheduling can be done using workload and cost estimates a day in advance, but the coincident peak is not known until the end of the month. Similarly, warnings that the next hour could be a coincident peak may only arrive from the utility with, in many cases, 5 minutes notice. Given the uncertainty about the coincident peak hour, 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 to minimizes the expected cost given the estimates of the likelihood of a coincident peak or warning during each hour of the day. For the worst case scenario, we propose a robust optimization based algorithm, which is computationally efficient, and guarantees that the cost is within a small constant of the optimal cost of an offline algorithm.

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**. Our experimental results 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 crucial to ensuring minimal energy costs and emissions. Specifically, combining workload shifting with local generation can provide 35-40% reductions of energy




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

(c) Energy costs (d) Emissions

**Figure 2: Comparison of energy costs and emissions for a data center with both local PV installations and local diesel generators. (a) and (b) show the plans computed by our algorithms.**

costs, and 10-15% reductions of emissions. An example of these results is shown in Figure 2, where we compare energy costs and emissions of our algorithms (termed "*Prediction (Pred)*" and "*Robust*", respectively) with two baselines meant to mimic current industry standard planning: *Night* tries to run jobs during night if possible and otherwise run these jobs with a constant rate to finish them before their deadlines, while *Best Effort* finishes jobs in a first-come-first-serve manner as fast as possible. *Optimal* is the offline optimal plan given knowledge of when the coincident peak will occur. As shown in the figures, our algorithms provide about 20% savings compared to *Night* and *Best Effort* (up to 40% in other cases). 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 2(a), while *Robust* tries to make the grid power usage as flat as possible as shown in Figures 2(b). 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. Our sensitivity analysis shows the costs and emissions of *Robust* are unaffected by the quality of the predictions; however the costs and emissions of *Prediction* change dramatically.

**Please refer to [1] for the full version.**

## Acknowledgements

This work was supported by NSF grants CNS 0846025, DoE grant DE-EE0002890, and HP Labs.

## 3. REFERENCES

[1] Z. Liu, A. Wierman, Y. Chen, B. Razon, and N. Chen, "Data center demand response: Avoiding the coincident peak via workload shifting and local generation [technical report],"
[2] National Institute of Standards and Technology, "NIST framework and roadmap for smart grid interoperability standards." NIST Special Publication 1108, 2010.
[3] Department of Energy, "The smart grid: An introduction."
[4] Federal Energy Regulatory Commission, "National assessment of demand response potential." 2009.
[5] NY Times, "Power, Pollution and the Internet."
[6] G. Ghatikar, V. Ganti, N. Matson, and M. Piette, "Demand response opportunities and enabling technologies for data centers: Findings from field studies," 2012.
[7] J. Koomey, "Growth in data center electricity use 2005 to 2010," *Oakland, CA: Analytics Press. August*, vol. 1, p. 2010, 2011.