Demand Response for Reducing Coincident Peak Loads in Data Centers

Maciej Z. Lukawski
Cornell University
mzl8@cornell.edu

Jefferson W. Tester
Cornell University
jwt54@cornell.edu

Michal C. Moore
Cornell University
mcm337@cornell.edu

Pawel Krol
AGH University of Science and Technology
pakrol@agh.edu.pl

C. Lindsay Anderson
Cornell University
cla28@cornell.edu

Abstract

Demand response is a key aspect of managing uncertainty and reducing peak loads in electric grids. This paper considers the capability of a datacenter to provide responsiveness to grid signals through cooling system control. The strategy is based on pre-cooling the center for provision of load reduction during demand response events, and is evaluated using a numerical model of a cooling system, validated against experimental data obtained from a small telecommunication data center. The pre-cooling strategy is applicable to a wide-range of demand response programs, but is illustrated on the example of an established critical peak pricing program; specifically the 4 coincident peak (4CP) program in the ERCOT ISO. Precooling reduced the annual cost of electricity used by the cooling system by 7.8 % to 8.6 %, while increasing the total energy use only by 0.05%. This translated into 2 % to 2.6 % reduction in the electric bill of the whole data center. The developed demand response strategy is suitable for data centers with power densities below 500 W/m² which do not use server air containment systems.

Nomenclature

a – thermal constant [-]
C – thermal capacitance [J/ºC]
COP – coefficient of performance for cooling (equal to cooling capacity divided by work input) [-]
h – time step [s]
m – constant determining whether cooling system is on (m = 1) or off (m = 0) [-]
n – number of active heat pump cooling stages: n = 1 for part-load and n = 2 for full-load [-]
Q_{IT} – internal heat gain from the IT equipment [W]
Q_C – cooling power of the heat pump [W]
R – overall thermal resistance [ºC/W]
t – time [s]
W – power consumption of the cooling equipment [W]
γ – ratio of the nominal cooling power to the IT load ($\gamma = Q_{C,nom} \cdot n/Q_{IT}$) [-]
θ – indoor temperature [ºC]
θ_a – ambient temperature [ºC]
θ_{min} – minimum thermostat set point [ºC]
$\theta_{max,1}$ – maximum thermostat set point (for part-load cooling) [ºC]
$\theta_{max,2}$ – maximum thermostat set point (for full-load cooling) [ºC]

Subscripts

avg – average
dr – demand response
nom – nominal (corresponding to 26.7 ºC indoor dry bulb temperature and 50% relative humidity)
pc – precooling
t – time

1. Introduction

Currently, there are a significant number of generators in the U.S. power system that exist to serve peak loads, mostly corresponding to periods when cooling demands are very high. At these times, electricity generating resources are scarce and the peaking plants are required to cover electricity shortages. The outcome is high electricity prices and output from power plants that are relatively inefficient and have higher emission levels [1].

Reduction in the grid-wide peak electric load (also known as coincident peak) can be accomplished through demand response. Demand response programs incentivize customers to alter their electricity use at specific times to improve reliability of the grid, for example by reducing coincident peak load. This is typically accomplished by charging customers a higher rate for electricity used during coincident peak periods [2] or by making the transmission charge dependent on the consumer's load during these times [3]. The coincident peak charge

URI: https://hdl.handle.net/10125/59776
ISBN: 978-0-9981331-2-6

may constitute upwards of 20% of the electric bill, providing a strong incentive for customers to reduce their demand during these periods [4]. The definition of what constitutes customer's contribution to the coincident peak varies by the regional transmission organization (RTO) or Independent System Operator (ISO). For example, the 4 coincident peak (4CP) program in ERCOT ISO defines it as the average of customer's loads during 15-minute system-wide peaks in each of 4 months from June to September [3,5]. The 5CP program in PJM RTO considers customers' contributions to the top 5 peak hours occurring on separate weekdays from June to September [5,6]. While it is impossible to predict the timing of the coincident peak events with certainty, many system operators, utilities, and consulting companies provide forecasts and alerts for customers participating in these programs [3].

The characteristics of data centers make them particularly suitable for participation in demand response programs including coincident peak pricing. Data centers represent approximately 1.8% of the total U.S. electricity consumption [7]. They are large, centralized loads which can provide substantial flexibility to the grid either by shifting their workload, or by temporarily adjusting the operation of their cooling systems [2]. In the latter approach, data centers act as thermostatically controlled loads (TCLs) and can be used to enhance grid reliability by providing ancillary services [8], load following, or by participating in energy arbitrage [9]. Cooling systems in data centers offer numerous advantages over residential TCLs: their cooling load is relatively constant year-round, they can provide fast response due to being oversized for reliability, and they may allow larger indoor temperature fluctuations compared to comfort-oriented residential heating and cooling systems. The large capacity data center cooling systems compared to residential air conditioning would also reduce the costs of control systems and load integration required for demand response [9].

Despite their large potential, most data centers don't actively participate in demand response [4]. The primary reasons are low financial incentives for many types of demand response and the fear of sacrificing the reliability of the IT equipment. The risk-averse data center operators are also not willing to give up any control over the cooling system to load aggregator or the utility, which is required by some demand response strategies [9].

The most common demand response program currently available to data centers in coincident peak pricing [4]. Data center demand response under coincident peak pricing programs has been studied for workload shifting and using local power generation [2], but not for shifting data center cooling loads. The latter approach is evaluated in this paper.

1.1. Scope of this paper

The goal of this paper is to develop a demand response mechanism for reducing coincident peak loads in low- to medium-power density data centers and server rooms. Such facilities typically use 30% to 50% of their energy demand for cooling IT equipment [10]. The proposed approach relies on precooling the data center prior to demand response events and reducing the cooling load during the peak. In the case of coincident peak programs, events are typically forecasted several hours in advance. If a coincident peak event is likely to occur on a given day, the data center operator can lower the indoor temperature ahead of the expected peak – a process referred to as precooling. During the predicted period of coincident peak, the cooling system can either be switched off or its output can be reduced, while only maintaining air circulation needed to avoid hot spots in the computer room.

The proposed approach has several advantages over alternative demand response mechanisms, making it more likely to be adopted by the risk-averse data center operators. It does not require expensive energy storage devices, it is easy to automate, and does not require surrendering the control over the cooling system to a distribution system operator or a load aggregator. The indoor temperature is always maintained below the allowable maximum, which alleviates the risk of overheating the IT equipment. In addition, the proposed strategy does not require an advanced communication platform between the systems operator and the data center. Instead, it relies on coincident peak forecasts, which are widely available. For example, in the ERCOT ISO such forecasting services were offered by 13 retail electricity providers, 8 municipal utilities, and a number of consulting companies [3].

The main limitation of the proposed control strategy is that it is applicable primarily to low- to medium- power density data centers and server rooms without server air containment systems. In order to provide sufficiently high demand response times, the power density calculated based on the total area of the server room should be below approximately 500 W/m^2. Such low power densities are characteristic of computer rooms, but may also be encountered in telecommunications data centers and other facilities with low floor utilization [10,11]. Our analysis shows that facilities with significantly higher power density do not have sufficient thermal storage capacity to provide extended demand response times.

This analysis begins with Section 2 discussing the numerical model used to simulate thermal behavior of data centers. This model is introduced in Section 2.1 and its parameters are fitted to the data acquired from a small telecommunications data center in Section 2.2. The results of the simulations are validated using experimental data in Section 2.3. Section 3.1 describes the control algorithm used for reducing coincident peak load. This algorithm is evaluated using a case study of the 4CP program in ERCOT. The 4CP program is discussed in Sections 3.2 and 3.3. The results of the case study are presented in Section 4 and sensitivity analysis is provided in Section 4.1. Lastly, concluding remarks are included in Section 5.

2. Modeling approach

2.1. Numerical model of telecommunications data center

To simulate the thermal behavior of a data center building and its cooling system we used a discrete time model adopted from previous work [8,12,13]. The parameters of the model were fitted to the data collected from a geothermal heat pump cooling system installed in a small telecommunications data center in Ithaca, NY. The model represents the data center as a homogenous thermal mass, which temperature is controlled by the heat transfer through the building envelope, the internal heat gains from the IT equipment, and the operation of the cooling system. The thermal inertia of the building is described by a dimensionless parameter a calculated as a function of the time step h, thermal capacitance C (in J/ºC), and thermal resistance R (in ºC/W):

$$a = exp\left(-\frac{h}{C \cdot R}\right) \quad (1)$$

The indoor temperature θ at time $t+1$ is calculated using Equation 2:

$$\theta_{t+1} = a \cdot \theta_t + (1-a)\left[\theta_{a,t} + R\left(Q_{IT} - m_t Q_{C,t}\right)\right] \quad (2)$$

Where θ_a is the ambient temperature (in ºC), Q_{IT} is the constant internal heat gain form the IT equipment (in W), and Q_C is the cooling power of the heat pump (in W). The binary constant m determines whether the cooling system is on or off depending on the minimum and maximum indoor temperature set points θ_{min} and $\theta_{max,1}$. The difference between $\theta_{max,1}$ and θ_{min} is the thermostat dead band.

$$m_{t+1} = \begin{cases} 0 \ if \ \theta_t \le \theta_{min} \\ 1 \ if \ \theta_t \ge \theta_{max,1} \\ m_t \ if \ \theta_{min} < \theta_t < \theta_{max,1} \end{cases} \quad (3)$$

The power consumption of the cooling system W (in W) is calculated based on its cooling capacity Q_C (in W) and the coefficient of performance COP:

$$W_t = \frac{1}{COP_t} \cdot Q_{C,t} \cdot m_t \quad (4)$$

The COP and the capacity of the cooling system Q_C are expressed as functions of the indoor temperature using Equations 5 and 6:

$$COP_t = COP_{nom} \cdot (0.022 \cdot \theta_t + 0.406) \quad (5)$$

$$Q_{C,t} = Q_{C,nom} \cdot n \cdot (0.024 \cdot \theta_t + 0.361) \quad (6)$$

where n is the number of active cooling stages of a heat pump (n = 1 for half-load and n = 2 for full-load). The second stage of the heat pump compressor (n = 2) is activated if the indoor temperature exceeds $\theta_{max,2}$ and deactivated if it drops below $\theta_{max,1}$. Both the nominal cooling capacity of the heat pump $Q_{c,nom}$ and the functional forms of Equations 5 and 6 were obtained from the heat pump specifications [14]. The overall coefficient of performance COP_{nom} at nominal conditions (i.e. 26.7 ºC indoor dry bulb temperature and 50% relative humidity) was obtained from a numerical model of the cooling system developed in TRNSYS software and validated using experimental data. The TRNSYS model was described in detail in our previous publication [15]. Equations 5 and 6 are applicable to geothermal cooling systems, in which heat sink temperature remains nearly constant throughout the year. The performance of conventional air-cooled computer room air conditioning (CRAC) systems would vary with the weather conditions, resulting in lower COP and cooling capacity at high ambient temperatures.

2.2. Fitting model to experimental data

2.2.1. Model calibration. An accurate depiction of the transient thermal behavior of the data center is essential for the validity of this study. For this reason, the parameters of the numerical model were estimated from the data collected from a small, 93 m² (1000 ft²) telecommunications data center located in Ithaca, NY. The data center had IT power demand of about 14 kW_e, indicating a low power density of 150 W/m² based on the computer room floor area. This is about one third of the median used power density in data centers worldwide [11]. The data center was equipped with a geothermal heat pump (GHP), which maintained the computer room temperature between θ_{min} of 25.6 and $\theta_{max,1}$ of 26.4 ºC (78 to 80 ºF). The second stage of the heat pump compressor was activated at $\theta_{max,2}$ of 27.2 ºC (81 ºF).

The schematic of the cooling system is shown in **Error! Reference source not found.**. The heat pump removes heat from the indoor air and transfers it through two hydronic loops connected by a heat exchanger to a series of 140 m deep borehole heat exchangers (BHEs). BHEs dissipate the heat to subsurface, providing a nearly constant temperature of liquid returning to the GHP independent of the weather conditions. The indoor air cooled by the GHP is distributed by a network of underfloor ducts and perforated tiles to the server room and returned through a system of vents located in the ceiling. The air flow is not contained to the IT equipment and the cold supply air is allowed to mix with the indoor air.

The geothermal heat pump provides a higher COP than air conditioning systems typically used in data centers and its performance is less sensitive to the ambient temperature. The dynamic thermal behavior of the data center, however, does not depend on the source of cooling and is expected to be the same for both geothermal and air-source systems.

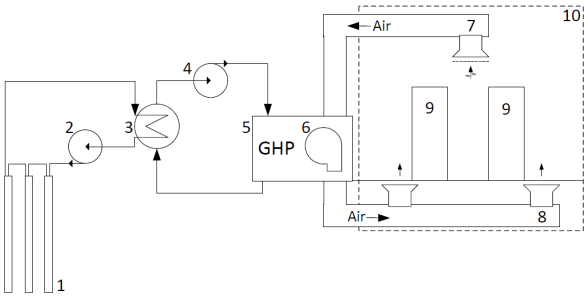

Figure 1: Schematic of the geothermal data center cooling system consisting of: 1) borehole heat exchangers, 2) wellfield glycol circulation pump, 3) glycol heat exchanger, 4) building circulation pump, 5) geothermal heat pump, 6) air blower, 7) hot air return, 8) cold air supply, 9) IT equipment, and 10) computer room.

The data center cooling system was equipped with a comprehensive monitoring and data acquisition system, which recorded relevant temperature, flow, and power consumption data in 5-minute intervals over a period of 43 days from August 15 to September 27, 2017. This data was used to estimate the values of parameters used in Equations 1 to 6, which were listed in Table 1. The nominal cooling capacity of heat pump per stage $Q_{C,nom}$ was calculated as an average over all times when the heat pump was on and the indoor temperature was within 1 ºC from 26.7 ºC. The heat pump can operate in two stages: at half-load (n = 1) or full-load (n = 2), with respective cooling capacities of approximately 15.8 kW and 31.6 kW. The electricity use of IT equipment was not directly measured, so the amount of heat generated by the IT equipment Q_{IT} was inferred from other measurements. The following procedure was used to calculate Q_{IT} and the thermal resistance R: for each time interval, the difference in the temperature of the ambient and indoor air ΔT was calculated. All recoded data were categorized into 2 ºC bins based on ΔT and the average cooling load for each bin was calculated. The data was binned to reduce the noise due to transient start up behavior of the heat pumps. The average cooling load was then plotted as a function of ΔT and a linear function was fitted to the data providing a coefficient of determination R^2 of 0.99. The Q_{IT} was set equal to the y-intercept of this function i.e. the cooling load for no heat transfer through the building envelope. The slope of the linear fit indicated the sensitivity of the cooling load to ΔT, and therefore R was calculated as the inverse of the slope. Lastly, thermal capacitance C was calculated by fitting the frequency of indoor temperature fluctuations from the model to experimental data for a period when the ambient temperature was stable and within 1 ºC from the indoor temperature. Due to negligible heat losses during this time, the heat load from equipment was set equal to the average cooling duty.

Table 1: Values of the model's parameters obtained from the experimental measurements.

Symbol	Parameter	Value
h	Time step	10 [s]
C	Thermal capacitance	$15.7 \cdot 10^6$ [J/ºC]
R	Thermal resistance	$4.67 \cdot 10^{-3}$ [ºC/W]
$Q_{C,nom}$	Nominal cooling capacity of heat pump per compressor stage	15767 [W]
Q_{IT}	Average heat gain from IT equipment	14040 [W]
θ_{min}	Minimum thermostat set point	25.6 [ºC]
$\theta_{max,1}$	Maximum thermostat set point for the part-load (n = 1) cooling	26.4 [ºC]
$\theta_{max,2}$	Maximum thermostat set point for the full-load (n = 2) cooling	27.2 [ºC]
COP_{nom}	Coefficient of performance at nominal conditions	3
N	Number of active heat pump compressor stages	1 or 2

2.2.2 Model validation. Figure 2 shows a comparison of the results from the numerical model to the data recorded at the telecommunications data center in Ithaca, NY. The dynamic thermal behavior simulated by the model is in good agreement with the data, as illustrated by similar period of fluctuations in the indoor temperature θ. Most importantly, the rate at which the indoor temperature increases when the cooling system is off is accurately captured by the model. The cumulative cooling load and electricity use are also in a good agreement, with approximately 2% and 4% difference between data and model results, respectively.

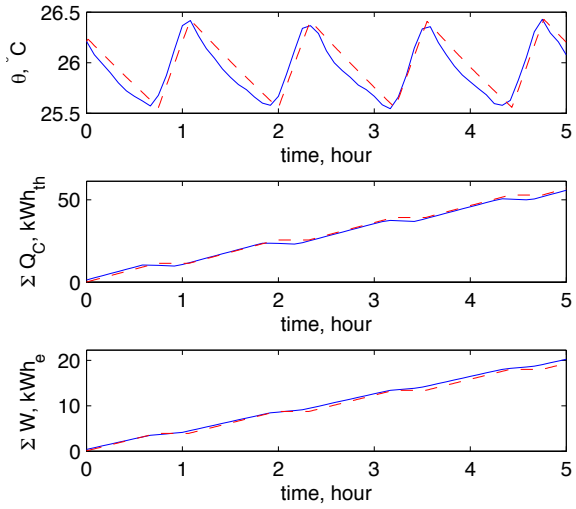

Figure 2: Validation of the results from the numerical model (red dashed line) with the experimental data (blue continuous line). Top: indoor temperature; middle: cumulative cooling duty; bottom: cumulative electricity used by the cooling system.

3. Approach to reducing coincident peak demand

This section discusses the proposed approach to reducing coincident peak load by precooling data centers or computer rooms ahead of the anticipated coincident peak events. The benefits and tradeoffs of this strategy are evaluated using an example of a telecommunications data center participating in the 4CP (4 Coincident Peak) program in ERCOT. This program serves as a useful example, given its relative maturity and accessible data.

3.1. System control

A schematic of the proposed demand response mechanism is presented in Figure 3. Both the predicted start time t_{dr} and the duration of the coincident peak warning Δt_{dr} are obtained by the data center using a forecasting service [5]. Based on the value of Δt_{dr}, the required indoor temperature to which data center needs to be precooled $\theta_{min,pc}$ is calculated using Equation 7:

$$\theta_{min,pc} = exp\left(\frac{\Delta t_{dr}}{C \cdot R}\right)\left[\theta_{max} + \left(exp\left(\frac{-\Delta t_{dr}}{C \cdot R}\right) - 1\right) \cdot \left(\theta_{a,dr} + R \cdot Q_{IT}\right)\right] \quad (7)$$

where $\theta_{a,dr}$ is the maximum dry bulb ambient temperature forecasted for the demand response period. If this information is not available, $\theta_{a,dr}$ can be conservatively estimated as the highest ambient temperature recorded in a given month during several recent years. In the example presented in Section 4, using a maximum monthly temperature from the last decade as $\theta_{a,dr}$ would lower the precooling temperature $\theta_{min,pc}$ only by 0.12 ºC and increase demand response time from 60 to 62 minutes.

Both the precooling temperature $\theta_{min,pc}$ and the rate of change in indoor temperature $d\theta/dt$ are subject to safety constraints. The ASHRAE A1 class has a *recommended* indoor temperature range of 18 to 27 ºC (64.4 to 80.6 ºF) and *allowable* range of 15 to 32 ºC (59 to 89.6 ºF) [16,17]. The ASHRAE A1 *allowable* class has a maximum indoor temperature change of 20 ºC in an hour [16], but some IT equipment vendors recommend $d\theta/dt$ below 5.5 ºC/hr [18].

The time needed to precool the data center Δt_{pc} can be then calculated using Equation 8:

$$\Delta t_{pc} = C \cdot R \cdot ln\left[\frac{\theta_t - \theta_{a,t} + R \cdot (Q_{C,avg} - Q_{IT})}{\theta_{min,pc} - \theta_{a,t} + R \cdot (Q_{C,avg} - Q_{IT})}\right] \quad (8)$$

where $Q_{C,avg}$ is the average cooling capacity during the precooling period:

$$Q_{C,avg} = 0.5 \cdot \left(Q_{C,t_{pc}} + Q_{C,t_{dr}}\right) \quad (9)$$

Figure 3 illustrates, that the precooling begins at time t_{pc}, which is Δt_{pc} before the beginning of the coincident peak alert at t_{dr}. If the indoor temperature $\theta_{min,pc}$ is achieved before t_{dr}, it is maintained at this low value until the beginning of the demand response. At time t_{dr} the cooling system is switched off and only the blower in the air handler is operated on a cyclic basis, as necessary to avoid hot spots in the computer room. This work assumed that the air blower would be switched on for 2 minutes for each 10 minute interval. The cooling system is activated again at the end of the coincident peak alert period $t_{dr} + \Delta t_{dr}$ or if the indoor temperature exceeds $\theta_{max,1}$.

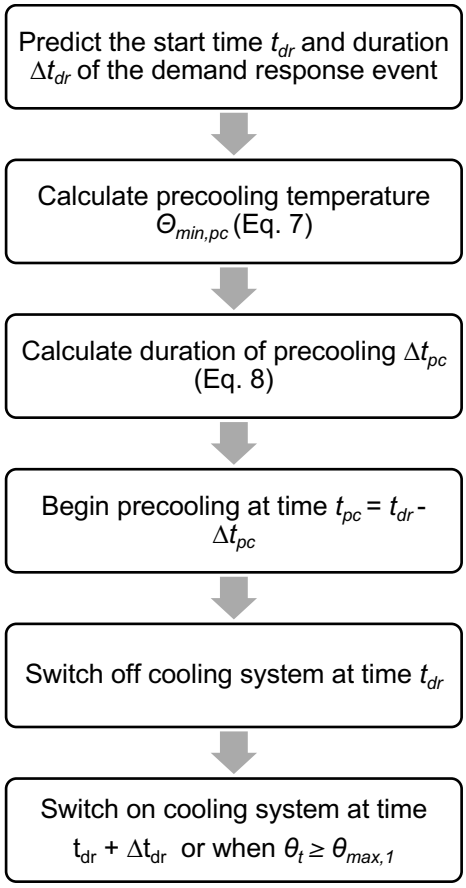

Figure 3: Schematic of the proposed demand response approach for reducing coincident peak demand in data centers

3.2. Case Study of Coincident Peak (4CP) program in ERCOT

The proposed demand response mechanism was evaluated using an example of the 4 Coincident Peak (4CP) program in ERCOT ISO. 4CP is a capacity charge program under which large commercial customers (>100 kW peak load) equipped with interval data recorders (IRDs) are charged transmission charge based on their average contributions to coincident peaks. The coincident peaks are four 15-minute periods with the highest grid-wide load, one during each of the months of June, July, August and September. The transmission charge is calculated based on the average consumer load during these four 15 min events in the previous calendar year [3,19].

The 4CP program was selected as a case study for this analysis because it is particularly suitable for the proposed demand response scheme. The 4CP alerts are infrequent, last a short time and can be accurately predicted due to the strong correlation between the ambient temperature and load. During the recent 6 years (2012 to 2017) all 4CP events in ERCOT occurred on weekdays between 4 PM and 5 PM, indicating that the required demand response time Δt_{dr} should not exceed 1 hour [19]. This compares favorably with the typical coincident peak alert durations of 2 to 6 hours issued in other markets [2,20]. The 4CP events can also be accurately predicted. A company offering 4CP forecasting services issued only 11 alerts in 2015, correctly predicting all 4 peak events [5]. So far, the customer response to the 4CP program did not result in significant peak shifting. Of the 28 4CP events between 2009 and 2015, only 5 were shifted by 15 minutes and 1 was shifted by one day due to demand response [21]. The high annual coincident peak charges of $18 to $23 per kW provide additional motivation for participation in the 4 CP program [22]. The installation of Interval Data Recorders needed for participation in the 4CP program is, however, required only for large consumers with non-coincident peak demand above 700 kW [3,23].

The challenge with coincident peak pricing programs, from the perspective of providing grid services, is in the limited number of events that can be called annually, which may become insufficient in particularly hot years, or unnecessary in cooler summer seasons, and does not take full advantage of the flexibility offered by data centers [4]. In addition, coincident peak programs may lead to peak-shifting under large scale deployment. However, for the case of large consumers, the 4CP program has been relatively effective for peak reduction, and provides a useful illustrative example for this strategy [3].

3.3. Methodology and assumptions used in 4CP case study

In this work, operation of a small telecommunications data center actively participating in the 4CP program was compared to the same data center not participating in demand response. The hypothetical data center was located in Houston, TX and the analysis was performed for the year 2017.

The thermal characteristics of the building and the cooling system were the same as in the experimental system Ithaca, NY. The nominal coefficient of performance of the system COP_{nom} was set at 2.7 and the nominal cooling capacity per heat pump per stage $Q_{C,nom}$ was set at 15 kW to reflect the higher heat sink temperature in Houston, TX compared to Ithaca, NY. Due to the higher ambient temperature, the heat pump operates more often on its full capacity rather than on

part-load. During precooling, only the main heat pump is used. Activating the backup cooling capacity would reduce the precooling time, but could also increase the non-coincident peak charge.

Two data center sizes were investigated:

a) 25 kW_e - a base case facility with a maximum combined load of approximately 25 kW_e and a constant IT load of 14 kW_e, identical to the experimental system installed in Ithaca, NY
b) 250 kW_e - a scaled up facility with 10 times greater maximum combined load

Under the current ERCOT regulations only the latter system would have a sufficient peak load (>100 kW) to qualify for the 4CP program, but this work assumed that both systems would be eligible for participation in 4CP.

To represent the accuracy and frequency of the 4CP alert services [5], this work assumed that the data center would need to respond to 11 peak warnings per year, each lasting 60 minutes, from 16:00 to 17:00. Four alerts predicted the actual 4CP events and the remaining seven were scheduled at non-4CP days with the highest ambient temperatures (2 warnings in June, July, and August and 1 in September).

The electricity prices for 2017 were obtained from two utilities participating in the 4CP program and are listed in Table 2. The combined monthly bills were calculated as sums of the transmission, distribution, and fixed charges determined by the utility [22], as well as real-time locational marginal prices for the Houston hub [24]. For simplicity, the transmission charge was determined based on the current, rather than previous year's 4CP load. The hourly ambient temperature records for Houston Intercontinental Airport were obtained from the National Oceanic and Atmospheric Administration [25].

Table 2: Generic monthly electricity cost components from two utilities for >10 kW customers equipped with Interval Data Recorders (ITD) and participating in the 4CP program. Calculations assume a power factor of one [22]. Energy charge is the locational marginal price for the Houston hub [24].

Monthly charges	CenterPoint Energy	AEP Texas
Customer charge ($ per customer)	65.83	26.52
Metering charge ($ per customer)	63.07	15.81
Transmission charge ($ per 4CP kW)	2.24	1.79
Distribution charge ($ per kW)	3.06	3.31
Energy charge ($ per kWh)	Real-time prices; avg. of 0.028	Real-time prices; avg. of 0.028

4. Results and discussion

Figure 4 illustrates the operation of the proposed demand response strategy during an actual 4CP event which occurred on July 28, 2017 from 16:45 to 17:00. The simulated response to a 4CP warning issued for the period from 4 to 5 pm (red dashed lines) was compared to the operation with no demand response (blue continuous lines). As a result of a heat gain through the building envelope and a lower cooling capacity in the hot climate, the heat pump remained on during the whole afternoon. In the system not participating in demand response, the heat pump oscillated between part-load and full-load from 15:00 to 19:30 to maintain the desired indoor temperature. The system participating in demand response began precooling at t_{pc} of 14:26, about 1.5 hour ahead of the 4CP warning period. During precooling, the indoor temperature was lowered to 22.5 ºC (72.5 ºF) at a rate of 2.8 ºC/hr, after which it increased at a rate of 3.9 ºC/hr during the 4CP warning period. Both the indoor temperature and its rate of change were well within industry safety standards [16–18]. During precooling, both the COP and the cooling output of the heat pump dropped as a result of the lower indoor temperature. The precooling strategy increased the electric consumption of the cooling system by 1.6% during the 24 hour period as a result of the lower COP and the cyclic operation of air blower.

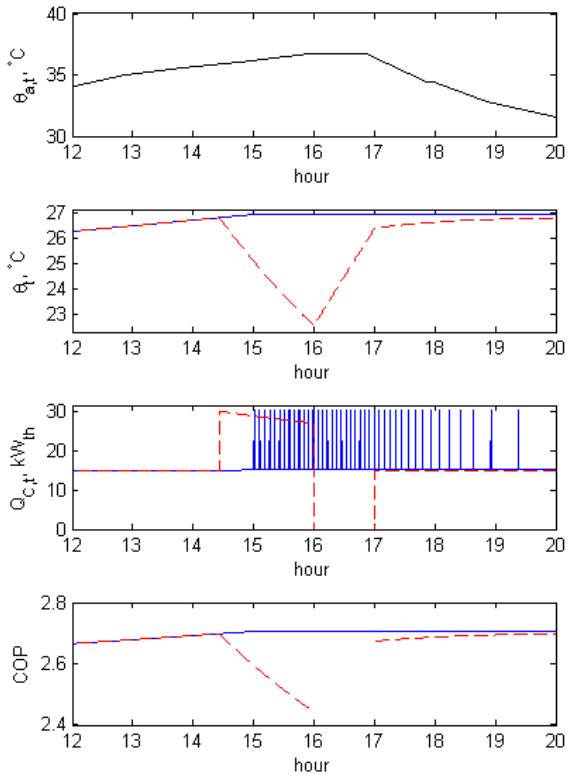

Figure 4: Simulated operation of a data center cooling system during a 4CP event on July 28, 2017. The operation with no demand response (blue continuous lines) is compared to the proposed precooling approach (red dashed lines). Figures (from the top): (1) ambient temperature $\theta_{a,t}$; (2) indoor temperature θ_t; (3) cooling output $Q_{C,t}$; (4) coefficient of performance COP.

As a next step, the whole year of operation of data centers with and without precooling was simulated. Demand response increased the total annual energy use of the cooling system by only 0.05 %, which is a very small tradeoff for increased load flexibility. The reduction in the energy bills, illustrated in **Figure 5** was much more significant. In the 25 kW_e data center, electric bills were reduced by approximately 2 %, which corresponds to $125 to $155 per year. In a 250 kW_e data center, the corresponding reduction was 2.1 % to 2.6 % ($1250 to $1550 per year). These savings correspond to 7.8 % to 8.6 % of the cost of electricity used by the cooling system, not accounting for the customer and metering charges. Such reduction is impressive given that the cooling system was deactivated for only 0.13 % of the total time. While a 2 % to 2.5 % reduction in the electric bill of a whole data center may not seem high, the approach may be profitable, particularly if automated and integrated with other demand response mechanisms. In addition, the cost reduction for traditional air-cooled CRAC systems would be about 50% higher than for geothermal heat pump (i.e. 3 % to 3.7 %) as a result of their lower COP [15].

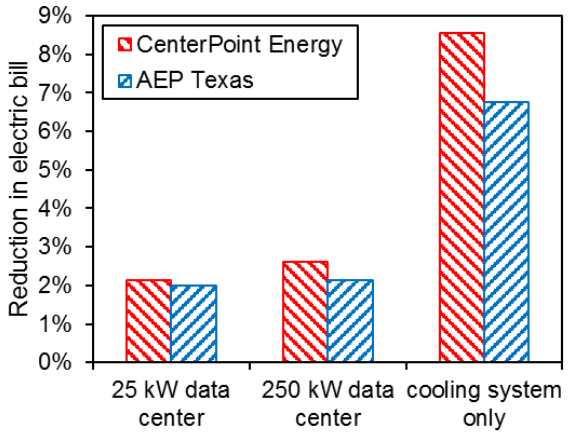

Figure 5: Percentage reduction in the electric bills due to precooling data centers prior to 4CP events.

4.1. Sensitivity to power density of the data center

The feasibility of precooling is largely dependent on the power density of data centers. The previous section discussed results for a data center with a computer room power density of 150 W/m^2. According to the 2011 survey of data centers, this value is below the median (437 W/m^2) and the 25th percentile (288 W/m^2) of the actual (used) power densities in data centers. This section evaluates precooling of data centers with power densities from 50 W/m^2 to 500 W/m^2 using example of the 4CP event on July 28, 2017. The precooling temperature $\theta_{min,pc}$, precooling time Δt_{pc}, and the maximum rate of change in indoor temperature $d\theta/dt$ were calculated by varying the IT load and cooling capacity and by assuming indoor temperature of 26 °C at the beginning of precooling.

Figure 6 shows the precooling temperature $\theta_{min,pc}$ and the maximum rate of change in indoor temperature $d\theta/dt$ as a function of the power density, calculated as the ratio of IT load to the computer room floor area. The $\theta_{min,pc}$ remains in the ASHRAE's *recommended* range of temperature (18 to 27 °C, continuous blue line in Figure 6) at power densities up to 350 W/m^2 and in the *allowable (A1)* range (15 to 32 °C, dashed blue line) up to 490 W/m^2. The maximum rate of change in indoor temperature $d\theta/dt$ exceeds 5.5 °C/hr recommended by some IT equipment vendors [18] for power densities above 230 W/m^2 (dashed red line in

Figure 6) but does not reach 20 ºC/hr allowed by the ASHRAE *allowable (A1)* class even at 500 W/m^2. Overall, precooling can provide one-hour demand response in data centers with power densities up to 490 W/m^2 and without server air containment systems if the ASHRAE's *allowable (A1)* class is adopted.

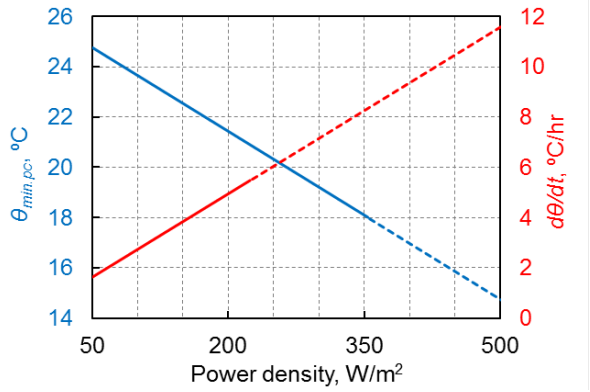

Figure 6: Sensitivity of the precooling temperature $\theta_{min,pc}$ and the maximum rate of change in indoor temperature $d\theta/dt$ to the power density of the computer room.

Figure 7 shows the precooling time Δt_{pc} presented as a function of the power density for various values of parameter γ, which is defined as the ratio of nominal cooling power used during precooling to the IT load. While most data centers have at least 100% backup capacity (corresponding to $\gamma = 2$), the use of excessive cooling capacity may increase the non-coincident peak load of the facility. Even a small amount of cooling capacity (γ of 1.1 to 1.2) available beyond what is needed to balance the heat generated by the IT equipment is sufficient to provide acceptable precooling times, typically below 2 hours. For a given value of γ, an increase in power density reduces the precooling time, as the impact of the heat gain through the building envelope becomes less meaningful.

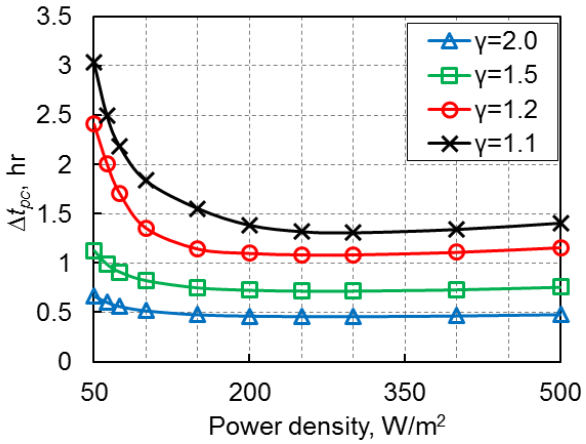

Figure 7: Sensitivity of the precooling time to the power density of the computer room. Parameter γ is the ratio of the nominal cooling power used during precooling to the IT load ($\gamma = Q_{C,nom} \cdot n / Q_{IT}$)

4.2. Limitations of this study and consideration for future work

Precooling proved to be a simple and cost effective way of reducing coincident peak charges in data centers with low- to medium-power density (<500 W/m^2), but this approach is less applicable to high density facilities. High-power density data centers often use air containment systems, which thermally isolate the IT equipment from the rest of computer room and drastically reduce the effective thermal capacitance [26,27]. As a result, switching off the cooling in high power density facilities can locally increase the indoor temperature at rates as high as 5 ºC per minute, making the proposed demand response mechanism infeasible [27]. High power density data centers could more effectively reduce their coincident peak loads by load shifting, using chilled water storage systems, or by running backup generators.

In addition to reducing coincident peak loads, precooling can be implemented as a response to price fluctuations in a real-time market. Our analysis of the 2017 marginal locational prices for the Houston hub indicated, that 10% of the total annual energy charge was incurred during 22 hours for which the electricity price was the highest. More than half of this time, the high price of electricity was sustained for 45 minutes or less [24], indicating that data centers could reduce these costs by implementing the proposed precooling approach. Data center precooling can also offset the instabilities resulting from the intermittent electric output of wind and solar plants and reduce both the curtailment of renewable resources and the ramping requirements for dispatchable generation.

5. Conclusions

The proposed demand response strategy can be used to reduce the coincident peak load in data centers and computer rooms. It relies on precooling the data center prior to a forecasted coincident peak event and switching off the cooling system during the peak, while only maintaining air circulation needed to avoid hot spots. The proposed strategy was evaluated using a case study of a small telecommunication data center with a power density of 150 W/m^2 participating in the 4CP program in the ERCOT ISO. Provided a coincident peak warning issued at least 1.5 hours prior to a 4CP event, the cooling system could deliver 1 hour long demand response, sufficient to avoid coincident peak charges in ERCOT. The proposed strategy provided a 7.8 % to 8.6 % reduction in the cost of electricity used by the cooling system, which corresponded to 2 % to 2.6 % reduction in the total electric bill of the data center. As a result of precooling, the total annual energy use of the cooling system increased only by 0.05 %. The proposed demand response strategy can be used in data centers and computer rooms without server air containment systems and with power densities below 500 W/m^2. It does not require existence of any advanced communications platforms between the distribution system operator and the end-users, and it can be integrated with other demand response mechanisms. The use of a pre-cooling control strategy provides a promising approach to load flexibility for data centers, wherein operators maintain control of the center loads. This strategy is also applicable to other types of demand response programs, and may provide increased cost savings, and grid benefits, under program structures.

6. Acknowledgements

The authors are thankful for the support from the Atkinson Center for a Sustainable Future at Cornell University in the form of Academic Venture Fund grant. We would also like to express our gratitude to a number of individuals who provided advice on this work: David Zurmuhl and Luke Sendelbach (Cornell University), Charles Kellum and David Hampton (Verizon Communications), Arian Aghajanzadeh, Dale Sartor, and Steve Greenberg (LBNL).

7. References

[1] U.S. EPA. Emissions & Generation Resource Integrated Database (eGRID) 2018. https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid (accessed March 3, 2018).

[2] Liu Z, Wierman A, Chen Y, Razon B, Chen N. Data center demand response: Avoiding the coincident peak via workload shifting and local generation. Perform Eval 2013;70:770–91. doi:10.1016/j.peva.2013.08.014.

[3] Zarnikau J, Thal D. The response of large industrial energy consumers to four coincident peak (4CP) transmission charges in the Texas (ERCOT) market. Util Policy 2013;26:1–6. doi:10.1016/j.jup.2013.04.004.

[4] Liu Z, Liu I, Low S, Wierman A. Pricing Data Center Demand Response. Acm Sigmetrics 2014:111–23. doi:10.1145/2591971.2592004.

[5] Genscape. Managing Capacity Charges with Genscape PowerBuyer 2015. https://www.genscape.com/blog/managing-capacity-charges-genscape-powerbuyer™ (accessed February 23, 2018).

[6] PJM. PJM Manual 19: Load Forecasting and Analysis. 2017.

[7] Shehabi A, Smith S, Sartor D, Brown R, Herrlin M, Koomey J, et al. United States Data Center Energy Usage Report. Berkeley, CA, USA. LBNL-1005775: 2016.

[8] Koch S, Mathieu JL, Callaway DS. Modeling and Control of Aggregated Heterogeneous Thermostatically Controlled Loads for Ancillary Services. Proc 17th Power Syst Comput Conf 2011:1–8. doi:10.1109/DICTA.2007.79.

[9] Mathieu JL, Dyson M, Callaway DS. Using Residential Electric Loads for Fast Demand Response□: The Potential Resource and Revenues, the Costs, and Policy Recommendations. Proc ACEEE Summer Study Build 2012:189–203.

[10] Cheung I, Greenberg S, Mahdavi R, Brown R, Tschudi W. Energy Efficiency in Small Server Rooms□: Field Surveys and Findings. 2014.

[11] Uptime Institute. Uptime Institute Annual Report□: Data Center Density. 2011.

[12] Callaway DS. Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy. Energy Convers Manag 2009;50:1389–400. doi:10.1016/j.enconman.2008.12.012.

[13] Mortensen RE. A stochastic computer model for heating and cooling loads. IEEE Trans Power Syst 1988;3:1213–9.

[14] ClimateMaster. Tranquility Compact Belt Drive (TC) Series Submittal Data. 2017.

[15] Zurmuhl DP, Lukawski MZ, Aguirre GA, Schnaars GP, Beckers KF, Anderson CL, et al. Hybrid Geothermal Heat Pumps for Cooling Telecommunications Data Centers. Proc. 43rd Work. Geotherm. Reserv. Eng. Stanford Univ., Stanford, CA, USA: 2018, p. 1–11.

[16] ASHRAE. Data Center Power Equipment Thermal

Guidelines and Best Practices. 2016.
[17] ASHRAE. ANSI/ASHRAE Standard 90.4-2016: Energy Standard for Data Centers. 2016.
[18] Oracle. Chapter 2: Environmental Requirements. Site Plan. Guid. Sun Servers, 2006.
[19] ERCOT. ERCOT Four Coincident Peak Calculations 2018. http://www.ercot.com/mktinfo/data_agg/4cp (accessed February 22, 2018).
[20] Northern Electric Cooperative. Understanding Demand & the Monthly Coincident Billing Peak 2018. https://www.northernelectric.coop/demand (accessed March 8, 2018).
[21] Raish C. Analysis of Load Reductions Associated with 4-CP Transmission Charges in ERCOT. Demand Side Work Gr Present 2016. http://www.ercot.com/content/wcm/key_documents_lists/87090/DSWG_4CP_Analysis_Raish.pptx (accessed February 23, 2017).
[22] Public Utility Commission of Texas. Comparison of Utilities' Generic Generic Transmission and Distribution Rates. Updated: September 1, 2017 2017. http://www.puc.texas.gov/industry/electric/rates/Trans/TDGeneric%0ARateSummary.pdf, (accessed February 15, 2018).
[23] Raish C, Turns L. ERCOT Impact Analysis of IDR Threshold Requirements. 2004.
[24] ERCOT. Historical real time marginal load zone and hub prices 2018. http://www.ercot.com/mktinfo/prices (accessed March 1, 2018).
[25] NOAA. Local Climatological Data, National Oceanic and Atmospheric Administration 2017. https://www.ncdc.noaa.gov/cdo-web/datatools/lcd (accessed February 15, 2018).
[26] Rasmussen N. The Different Types of Air Distribution for IT Environments. 2017.
[27] Lin P, Zhang S, VanGilder J. Data Center Temperature Rise During a Cooling System Outage. 2014.