Ramping Rate Flexibility of Residential HVAC Loads

Borhan M. Sanandaji[b], Tyrone L. Vincent[c], and Kameshwar Poolla[b]

***Abstract*— Residential Air Conditioners (ACs), refrigeration units, and forced air heating loads are candidates for providing ancillary services to the grid. Motivated by the need for resources with high ramping rate capability, we investigate the ramping rate flexibility of such loads and show that a collection of residential Heating, Ventilation, and Air Conditioning (HVAC) loads can provide regulating reserve service with certain ramping rate bound that is a result of enforcing a no-short-cycling requirement. A load is called short-cycled if it is switched ON and OFF quicker than a certain allowed time. We support our proposed bounds and theorems with illustrative simulations.**

I. Nomenclature

C	Thermal capacitance [kWh/°C]
R	Thermal resistance [°C/kW]
P_m	Rated electrical power [kW]
η	Coefficient of performance
θ_r	Temperature set-point [°C]
Δ	Temperature deadband [°C]
θ_a	Ambient temperature [°C]
$\theta^k(t)$	Temperature of unit k at time t
P_m^k	Power draw of unit k when it is ON
P_{tot}	$\sum_k P_m^k$
P_a^k	Average power consumed over a cycle
$n(t)$	$\sum_k P_a^k$, Baseline power consumption of the collection
P_o^k	Nominal power required to keep unit k at its set-point
P_{ave}	$\sum_k P_o^k$
$P_{\text{agg}}(t)$	$\sum_k q^k(t) P_m^k$, Instantaneous power drawn by units
$r(t)$	Regulation signal
$P^{\text{lim}}_{\text{ON}\to\text{OFF}}(t)$	Total power of units switched from ON to OFF at time t due to temperature bounds
$P^{\text{lim}}_{\text{OFF}\to\text{ON}}(t)$	Total power of units switched from OFF to ON at time t due to temperature bounds
$P_{\text{ON}}(t)$	Total power of ON units at time t
$P_{\text{OFF}}(t)$	Total power of OFF units at time t
$P^{\text{avail}}_{\text{ON}}(t)$	Total power of units that are available ON
$P^{\text{unavail}}_{\text{ON}}(t)$	Total power of units that are unavailable ON
$P^{\text{avail}}_{\text{OFF}}(t)$	Total power of units that are available OFF
$P^{\text{unavail}}_{\text{OFF}}(t)$	Total power of units that are unavailable OFF

[b]Borhan M. Sanandaji and Kameshwar Poolla are with the EECS department, University of California, Berkeley, CA 94720. Corresponding author's email: sanandaji@eecs.berkeley.edu.

[c]Tyrone L. Vincent is with the Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401.

Supported in part by EPRI and CERTS under sub-award 09-206; PSERC S-52; NSF under Grants CNS-0931748, EECS-1129061, CPS-1239178, and CNS-1239274; the Republic of Singapore National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore for the SinBerBEST Program; Robert Bosch LLC through its Bosch Energy Research Network funding program.

II. Introduction

DEEP penetration of Variable Energy Resourcess (VERs) (mainly wind and solar generation) into the current electric grid is necessary yet challenging to ensure grid reliability. Integration of such resources creates new operating grid conditions. To illustrate these new conditions, California Independent System Operator (CAISO) has created future scenarios of "net load" curves where net load is defined as forecasted load minus predicted variable generation. The so-called "duck curve" is an example of such studies. Fig. 1 depicts a net load curve for March 31 for years 2012 through 2020. As can be seen, a decline in mid-afternoon followed by a high ramp creates a curve that is similar to the neck of a duck. In particular, the most significant daily ramp starts around 5:00 p.m. when the sun sets (i.e., solar generation ends) and the demand increases. This requires the system operator to provide resources with ramping rate flexibility of 12,000 MW over 3 hours (around 67 MW/min). A ramp rate is the rate, expressed in MW per minute, at which a resource can change (increase or decrease) its output generation. Most industrial gas turbines, for instance, have an average ramp rate of 25 MW/min, indicating the incapability of such thermal generators in achieving the required ramp rate flexibility. As a result, these studies have forced regulatory agencies to propose new orders for operating the gird on shorter time frames and to motivate flexible resources with high up and down ramping capabilities. For example, a recent Federal Energy Regulatory Commission (FERC) order (Order No. 764) proposes to use *intra-hour* transmission scheduling (at 15-minute intervals) instead of existing hourly scheduling protocols. The order argues that intra-hour scheduling would increase transmission system flexibility and efficiency, providing grid operators with more options for scheduling resources during each hour and decreasing the need for (and costs of) ancillary services needed for reliable integration of VERs [1].

Frequency regulation is one of the most important ancillary services for maintaining the power balance between supply and demand in normal grid operating conditions [2]. It is deployed in seconds (up to one minute) time scales to compensate for short term fluctuations in the net load. This service has been traditionally provided by either fast responding generators or grid-scale energy storage units. However, the current storage technologies have high cost while generation has both cost and an environmental footprint. Moreover, traditional generators have slow ramping rates and cannot track a fast changing regulation signal very well [3]. On the other hand, increased penetration of renewable energies results in higher regulation requirements on the grid [4]–[6]. For instance, it has been shown that if California adopts its 33% Renewable Portfolio Standard (RPS) by 2020, the regulation procurement is an-

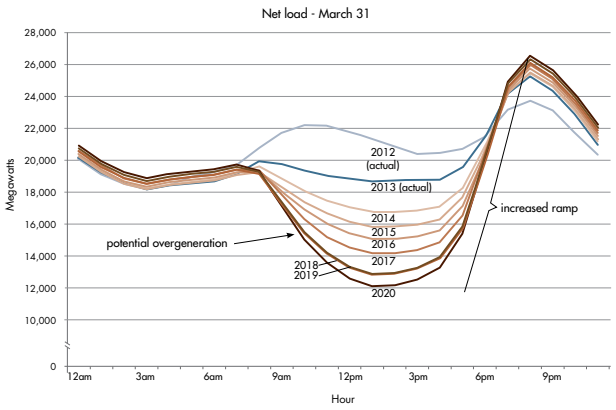


Fig. 1: CAISO's duck curve showing steep ramping needs [13].

ticipated to increase from 0.6 GW to 1.4 GW [7], [8]. Such requirements can be lowered if *faster* responding resources are available [9]. For instance, it has been shown that if CAISO dispatches fast responding regulation resources, it would reduce its regulation procurement by 40% [10]. These factors coupled with the search for cleaner sources of flexibility as well as regulatory developments such as FERC order 755 (2011) [11] and 784 (2013) [12] have motivated a growing interest in exploiting fast responding *demand-side* resources for enabling deep renewable integration.

Flexible loads including Electric Vehicles (EVs), residential, and commercial buildings have been recently considered as good candidates for providing ancillary services to the grid [3], [14]–[33]. Of our particular interest are residential Thermostatically Controlled Loads (TCLs) which require motors/compressors for their function such as Air Conditioners (ACs), refrigeration units or forced air heating loads [14]–[17], [19], [34]. These types of electric loads represent about 20% of the total electricity consumption in the United States [35], [36], and thus present a large potential for providing various ancillary services to the grid. Leveraging the inherent thermal storage of TCLs, their electricity consumption can be varied while still meeting the desired comfort level and temperature requirements of the end user. Recently, it has been shown that the aggregate flexibility offered by a collection of residential TCLs can be succinctly approximated using Generalized Battery Models (GBMs) with dissipation [37].

The main contribution of this paper is to derive bounds on the ramping rate flexibility of residential ACs, refrigeration units or forced air heating loads in providing frequency regulation. In particular, we show that a collection of such loads has certain ramping rate bound that is a result of enforcing a no-short-cycling requirement. In order to avoid wear and tear, TCL manufacturers require a minimum duration of time between any two consecutive ON/OFF switchings. If this minimum time is not met, the unit is said to be short-cycled. Consequently, a characterization of regulation signals that can be feasibly met by an aggregation of such loads is the intersection of this new constraint on the first difference of the regulation signal and signals feasible for the GBM. We provide a brief summary of this model in Appendix A of this paper. The proposed GBM resembles a model for physical storage device with components to represent the power drawn from/supplied to a battery as well as its State of Charge (SoC).

The remainder of the paper is organized as follows. Section III briefly describes the considered individual residential TCL model. We derive ramping rate constraint to address the no-short-cycling requirement of such loads in Section IV. We present our simulation results in Section V and conclude in Section VI. To be more accurate, we would call the considered flexible loads in this paper as residential Heating, Ventilating, and Air Conditioning (HVAC) loads to distinguish from TCLs such as electric water heaters which do not require motors/compressors to function. However, we interchangeably use TCLs and residential HVAC loads throughout the paper as they share the same models and characteristics.

III. Individual TCL Model

Let $\theta^k(t)$ be the internal temperature of the kth TCL at time t, θ_a be the ambient temperature, and P_m^k be its rated electrical power. Then, the temperature evolution of the kth TCL can be described by a standard deadband model as

$$\dot{\theta}^k(t) = \begin{cases} -a^k(\theta^k(t) - \theta_a) - b^k P_m^k + w^k(t), & \text{ON state}, \\ -a^k(\theta^k(t) - \theta_a) + w^k(t), & \text{OFF state}, \end{cases} \quad (1)$$

where $a^k := 1/C^k R^k, b^k := \eta^k / C^k$, and R^k, C^k, and η^k are model parameters as described in Table I. For more details on individual TCL models, please see [14], [16], [38].[1] Each TCL has a temperature set-point θ_r^k with a local ON/OFF control within a deadband $[\theta_r^k - \Delta^k, \theta_r^k + \Delta^k]$. The average power consumed by the kth TCL over a cycle is

$$P_a^k = \frac{P_m^k T_{\text{ON}}^k}{T_{\text{ON}}^k + T_{\text{OFF}}^k},$$

where T_{ON}^k and T_{OFF}^k are given by

$$T_{\text{ON}}^k = R^k C^k \ln \frac{\theta_r^k + \Delta^k - \theta_a + R^k P_m^k \eta^k}{\theta_r^k - \Delta^k - \theta_a + R^k P_m^k \eta^k},$$
$$T_{\text{OFF}}^k = R^k C^k \ln \frac{\theta_r^k - \Delta^k - \theta_a}{\theta_r^k + \Delta^k - \theta_a},$$

and represent the ON and OFF state durations per cycle, respectively. The baseline power consumption of k units is

$$n(t) := \sum_k P_a^k.$$

The aggregated instantaneous power consumption is

$$P_{\text{agg}}(t) := \sum_k q^k(t) P_m^k,$$

where $q^k(t) = 1$ when the TCL is ON and $q^k(t) = 0$ when it is OFF. In [3], [38], we conjectured that the aggregate behavior of a population of TCLs under the deadband model (1) can be accurately represented using a continuous-power model as

$$\dot{\theta}^k(t) = -a^k(\theta^k(t) - \theta_a) - b^k p^k(t), \quad (2)$$

[1] As common in the literature, $w^k(t)$ in model (1) is assumed to be Gaussian with zero mean and small variance [14], [16], [39], and can be neglected.

TABLE I: Typical parameter values for a residential AC.

Parameter	Description	Value	Unit
C	thermal capacitance	2	kWh/°C
R	thermal resistance	2	°C/kW
P_m	rated electrical power	5.6	kW
η	coefficient of performance	2.5	
θ_r	temperature set-point	22.5	°C
Δ	temperature deadband	0.3125	°C
θ_a	ambient temperature	32	°C

where $p^k(t) \in [0, P_m^k]$ is a continuous-power input to each TCL. In order to further legitimize the use of the continuous-power model in the analysis, in the following, we actually prove this conjecture for a homogenous collection of TCLs.

Theorem 1: Consider a homogeneous collection of N TCLs, each with parameters $(a, b, \theta_a, \theta_r, \Delta, P_m)$. A continuous-power model is described as

$$\dot{\theta}_c(t) = -a(\theta_c(t) - \theta_a) - bP_m p(t), \tag{3}$$

where $p(t) \in [0, 1]$. Assume the initial conditions satisfy $\frac{1}{N}\sum_k \theta^k(0) = \theta_r$ and $\theta_c(0) = \theta_r$. Then,

1) Given binary sequence $q^k(t)$ such that $\theta^k(t) \in [\theta_r - \Delta, \theta_r + \Delta]$, then by applying $p(t) = \frac{1}{N}\sum_k q^k(t)$ to the model (3) we have $\theta_c(t) \in [\theta_r - \Delta, \theta_r + \Delta]$ for all $t > 0$.

2) Assume $p(t)$ to be step-wise constant over time T, where T is small enough such that $P_m b(1 - e^{-aT})/a < \epsilon$. Then for any $p(t)$ meeting the temperature bounds of the continuous-power model, there exists binary sequence $q^k(t)$ such that $\|p(t) - \frac{1}{N}\sum_k q^k(t)\|_\infty \leq \frac{1}{N}$ and $\left|\theta^k(t) - \theta_r\right| \leq \Delta(1 + 2e^{-at}) + \epsilon\left(1 + \frac{1}{N}\right)$.

Proof: See Appendix B. ■

Theorem 1 describes the equivalence between the deadband model (1) and the continuous-power model (2) for a collection of homogeneous TCLs. One would like to assume that if $p(t)$ is a feasible power signal for the continuous-power model, then there exists binary sequence $q^k(t)$ such that the deadband model has close to the same average power draw, while each unit does not exceed the temperature limits. However, there are some fundamental issues if $p(t)$ is at the lower or upper temperature boundaries and the deadband model temperatures θ^k have a wide distribution. For example, if $p(t) = 1$, which requires all units to be ON at time t, and we have many units at time $t = 0$ with $\theta^k = \theta_r^k - \Delta^k$, then we cannot achieve both close tracking of $p(t)$ and satisfy the temperature limits. However, this is an issue simply because the system has no control over the initial temperatures. Thus, our results show that we can closely approximate $p(t)$ with an error that decreases linearly in N, while meeting temperature constraints that converge asymptotically to the desired temperature constraints within an arbitrary tolerance. The rate of convergence in this proof depends on the parameter a, which is an effect of assuming the worst case $p(t)$ and initial distribution of θ^k. In the remainder of this paper, we use the continuous-power model for analysis but use the deadband model in simulations. Extending Theorem 1 to the case of heterogenous collection is a much harder task and is under our current investigation.

Under the continuous-power model (2), a nominal power

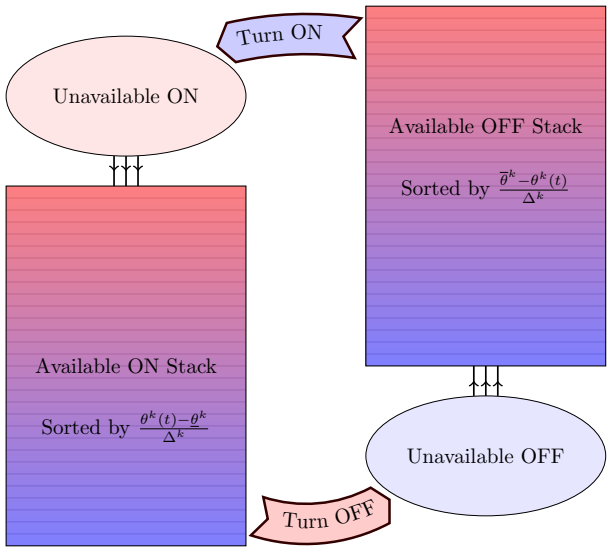


Fig. 2: Augmented ON and OFF stacks considering no-short-cycling constraints. A warmer unit has a higher priority to be switched ON and a cooler unit has a higher priority to be turned OFF. However, when no-short-cycling constraints are imposed, we are only allowed to manipulate units that are Available ON (Definition 2) or Available OFF (Definition 1). The lower and upper temperature bounds are given by $\underline{\theta}^k = \theta_r^k - \Delta^k$ and $\overline{\theta}^k = \theta_r^k + \Delta^k$.

required to keep the kth TCL at its set-point θ_r^k is

$$P_o^k = \frac{a^k(\theta_a - \theta_r^k)}{b^k} = \frac{\theta_a - \theta_r^k}{\eta^k R^k}. \tag{4}$$

Note that P_o^k is a random variable that depends on the ambient temperature and user-defined set-points. Rather trivial calculations reveal that the nominal power P_o^k under the continuous-power model closely follows the average power P_a^k under the deadband model for a range of operating conditions.

IV. Ramp Rate Constraints on Regulation Signal

After introducing the individual TCL model, we now present our *augmented* priority-stack-based control framework for manipulating the power consumption of a population of TCLs under a no-short-cycling constraint. Moreover, we analytically characterize the no-short-cycling constraint in terms of bounds on the ramping rate of the regulation signal.

A. Priority-Stack-Based Control

A centralized approach: We adopt a centralized control architecture. In our view, a purely decentralized control strategy is very difficult to realize. Some feedback signal available locally is required to assure that the correct amount of ramping is provided. While accurate and inexpensive local frequency measurements are available, these are not enough for flexible loads to deduce the ramping signal. Latency in the control loop is another difficulty. Distributed schemes with lower communication burden are possible, and there is some literature exploring message passing and consensus algorithms

for frequency regulation which may be adapted to ramping. Our view is that centralized control schemes offer much better reliability which is an enormous concern in power systems operations. Auditing and verification of services is also much more transparent here. The communications and infrastructure burden is not significantly higher than with a distributed control architecture. This choice is also dictated by the stringent power quality, auditing and telemetry requirements necessary to participate in regulation service market [40], [41].

Control Framework: At each sample time, the aggregator compares the regulation signal $r(t)$ with the aggregate power deviation $\delta(t) = P_{\text{agg}}(t) - n(t)$, where $P_{\text{agg}}(t)$ is the instantaneous power drawn by TCLs and $n(t)$ is their baseline power.

If $r(t) < \delta(t)$, the population of TCLs needs to "discharge" power to the grid which requires turning OFF some of the ON units. Conversely, if $r(t) > \delta(t)$, then the population of TCLs must consume more power. This requires turning ON some of the OFF units. To track a regulation signal $r(t)$, the system operator needs to determine appropriate switching actions for each TCL so that the power deviation achieved by collection of TCLs, $\delta(t)$, follows the regulation signal, $r(t)$.

It is reasonable to design a controller that first turns ON (OFF) warmer (cooler) units (giving them a higher priority). To this end, a priority-stack-based control method has been considered [18], [37], [42] in which units are sorted based on their temperature distance to the switching boundaries. An OFF unit that is closest (smallest temperature distance) to the upper temperature boundary turns ON first. The next OFF units with larger temperature distances to the upper boundary (lower priorities) turn ON in sequence until the desired regulation is realized. The priority-stack-based control strategy attempts to minimize the ON/OFF switching action for each unit.

B. No-Short-Cycling Requirement

The proposed priority-stack-based control scheme attempts to reduce the frequent switchings of each TCL. However, it can not guarantee that none of the units will not be switched quicker than allowed. To this end, one should *explicitly* impose such no-short-cycling constraints as part of the control strategy. Once a unit is turned ON or OFF, it must remain in that state for at least a certain amount of time (that is specified by the manufacture) before it is switched again. Some similar considerations have been taken into account in a concurrent work [43] where the number of times a load can be turned ON or OFF in a given period is studied. Another relevant work by Koutitas [20] considered user's comfort in its proposed binary ON/OFF control policies and further proposed round robin scheduling algorithms to sustain fairness in the system. Our work differentiates from these papers by analytically characterizing the ramping rate flexibility of the collection of loads in providing a regulation service when a no-short-cycling constraint is enforced by the control algorithm. In the following, we explicitly differentiate ON and OFF units based on their availability in participating in the control strategy.

Definition 1 (Unavailable (Available) OFF Units): Given τ_0, a unit is called unavailable (available) OFF when it has been OFF for less (more) than τ_0 since its last switching. $P_{\text{OFF}}^{\text{avail}}(t)$ is the total rated power of available OFF units. □

Definition 2 (Unavailable (Available) ON Units): Given τ_1, a unit is called unavailable (available) ON when it has been ON for less (more) than τ_1 since its last switching. $P_{\text{ON}}^{\text{avail}}(t)$ is the total rated power of available ON units. □

Fig. 2 depicts the proposed available ON and OFF priority stacks. When a unit is turned OFF (either due to normal control scheme or due to its participation in frequency regulation), it will remain unavailable for a certain amount of time before it is added to the available OFF stack. Similarly, when a unit is turned ON, it will be unavailable for a certain amount of time before it is added to the available ON stack. In both situations, units will be placed in the available stacks based on their temperature distances to the switching boundaries.

Unavailable ON and OFF units can not participate in regulation service provision. As mentioned earlier, when we require the control algorithm to explicitly satisfy the no-short-cycling requirement, a certain percentage of TCLs will be unavailable ON or OFF. The effect of this loss of use is to create an additional constraint on *changes* in feasible regulation signals $r(t)$. Quite simply, if there are no available ON units to be switched OFF, the regulation signal cannot request decreased power draw (and similarly for increased power draw). To determine feasible regulation signals, the battery model must be augmented with additional constraints on the ramping rate of the regulation signal. In what follows, we derive such bounds to prevent short cycling of TCLs. We first start by characterizing the total power of available ON and OFF units.

Theorem 2: Assume a collection of TCLs defined by P_m^k and P_o^k. Let $P_{\text{tot}} = \sum_k P_m^k$ and $P_{\text{ave}} = \sum_k P_o^k$. Let $\tau_0 > 0$ (Definition 1) and $\tau_1 > 0$ (Definition 2) be the minimum duration of time that a unit should remain OFF or ON, respectively. Let $P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t)$ and $P_{\text{ON}\to\text{OFF}}^{\text{lim}}(t)$ denote the total rated power of units that are about to be turned ON or OFF by local deadband control due to their temperature limits, respectively. If the regulation signal $r(t)$ is satisfied at t, then

$$P_{\text{OFF}}^{\text{avail}}(t) = P_{\text{tot}} - P_{\text{ave}} - r(t) - \sum_{k=t-\tau_0}^{t} \left(P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k) + [-D(k)]_+\right),$$

$$P_{\text{ON}}^{\text{avail}}(t) = P_{\text{ave}} + r(t) - \sum_{k=t-\tau_1}^{t} \left(P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) + [D(k)]_+\right),$$

where $D(t) := \Delta r(t) - (P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t) - P_{\text{ON}\to\text{OFF}}^{\text{lim}}(t))$, $\Delta r(t) := r(t) - r(t-1)$, and $[x]_+ := \max(x, 0)$.

Proof: See Appendix C. ■

Theorem 2 characterizes the total available power that can be provided by ON and OFF units, while the Automatic Generation Control (AGC) signal is met. We now show that how imposing the no-short-cycling requirement on TCLs results in a ramping rate constraint on the regulation signal.

Theorem 3: Let $P_{\text{OFF}}^{\text{avail}}(t)$ and $P_{\text{ON}}^{\text{avail}}(t)$ be as given in Theorem 2. Then, a regulation signal meets the no-short-cycling requirement if $\Delta r(t)$ is bounded as

$$-\mu_-(t) \le \Delta r(t) \le \mu_+(t), \tag{5}$$

where

$$\mu_+(t) = P_{\text{OFF}}^{\text{avail}}(t-1) - \max(P_{\text{ON}\to\text{OFF}}^{\text{lim}}(t), P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t)),$$

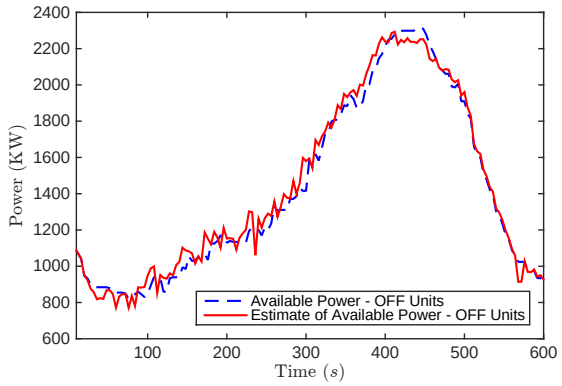


Fig. 3: Available power from OFF units and its estimation (Definition 1).

$$\mu_-(t) = P_{\text{ON}}^{\text{avail}}(t-1) - \max(P_{\text{ON}\to\text{OFF}}^{\text{lim}}(t), P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t)).$$

Proof: See Appendix D. ■

Remark 1: Theorem 3 provides *time-varying* bounds on $\Delta r(t)$ to guarantee no-short-cycling. Clearly, these bounds depend on the regulation signal itself, complicating an easy interpretation. However, there are important practical uses. Philosophically, this explicitly demonstrates an additional constraint beyond those implied in a battery model that a regulation signal should satisfy to avoid short cycling individual units. More practically, these bounds can be used to verify that a specific known regulation signal (or perhaps a family of regulation signals) could be sustained. This does require a simulation of the aggregation in order to determine the required signals (specifically, $P_{\text{ON}\to\text{OFF}}^{\text{lim}}(t)$ and $P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t)$), but since the system dynamics are quite simple, there are many options available, including simplified models such as Bin State Transition Models (BSTMs) [3], [16], [21], [26].

V. Simulation Results

In this section, we testify the proposed bounds on the ramping rate of the regulation signal. Consider a heterogeneous collection of 500 ACs whose model parameters are drawn from a uniform distribution with 10% heterogeneity around some nominal values (as listed in Table I). The proposed control scenario works with any heterogeneity level of TCLs and the chosen 10% heterogeneity level is just an example of a typical real-life situation. The priority-stack-based controller is applied to track a 10-minute long regulation signal $r(t)$ from Pennsylvania-New Jersey-Maryland (PJM) shown as the solid line in Fig. 5(a) [44]. The magnitude of the PJM signal is scaled appropriately to match the power limits and energy capacity of 500 ACs. In our simulations, we assumed that a unit is to remain in its current state for at least 1 minute after each switching. Fig. 4 illustrates the temperature profile of 2 random TCLs under the proposed control strategy. As can be seen, the 2 units have different temperature setpoints and boundaries due to the existing 10% heterogeneity of the collection. More importantly, the 2 units have quite different ON and OFF cycling behavior while satisfying the no-short-cycling constraint. Fig. 3 shows the available power from OFF units of this TCL collection derived by this regulation signal. Its estimate is also calculated based on Theorem 2 and is plotted. For comparison, the power and capacity limits found using the GBM are also shown in Fig. 5(a) and Fig. 5(b), respectively. As can be seen, the power and capacity limits are not violated by this regulation signal. We take $P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t)$ and $P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t)$ as that reported by the local unit controllers (They can be determined by knowing which units turn OFF (or ON) and their power draw), and use that information along with $r(t)$ to calculate $\mu_+(t)$ and $\mu_-(t)$. In Fig. 5(c) these are plotted along with $\Delta r(t)$. Around time 400 (s), the lower bound approaches zero, meaning that negative $\Delta r(t)$ is no longer feasible. Fig. 5(d) depicts the difference between the desired regulation signal and the actual power draw $P_{\text{agg}}(t) - P_{\text{ave}}$, confirming that regulation signal is not well followed during the time that $-\mu_-(t)$ is close or equal to zero.

VI. Conclusions

An aggregation of TCLs can be utilized to provide fast regulating reserve service for power grids and the behavior of the aggregation can be succinctly approximated using Generalized Battery Models (GBMs) with dissipation. In this paper, we analytically characterized the impact of imposing a no-short-cycling requirement on TCLs in terms of constraints on the ramping rate of the AGC signal. A characterization of regulation signals that can be feasibly met by the aggregation is the intersection of the constraints on the first difference of the regulation signal and signals feasible for the GBM. We believe that the proposed scheme has the modest computation and communication overhead. We have tried up to 10,000 TCLs in simulations without any degradation in performance. The computational cost of the proposed algorithm and the priority-stack-based control is fairly low. It basically includes a sorting step which using advanced sorting algorithms such as QuickSort or MergeSort has time complexity of $\mathcal{O}(N \log N)$ where N is the number of TCLs. Another concern is the communication burden of the proposed algorithm which includes receiving/reading TCL temperatures and set-points and sending out ON/OFF control commands. It is a reasonable assumption that nowadays such communication burden can be fairly handled with short delays. In our previous works, we studied such communication concerns and delays.

As a final note, in our view a realistic response to provision of ramping resources requires a mix of assets. Flexible loads such as TCLs have some key cost advantages but are less reliable than conventional generation. For example, ramping capacity from TCLs is likely very different of weekdays and weekends and is random because of occupancy fluctuations and weather. On the other hand, if conventional generation resources are used for infrequent (once/day) ramping, their capacity factor is low making them very expensive. Flexible loads do not suffer from "opportunity cost" if their flexibility is not used. We would imagine that a mix of assets would be the best practical solution. The bulk of ramping could be supplied by flexible loads, while fossil fuel generation could be sized to cover the uncertainty of ramping capacity supplied by flexible loads. Coordination of these resources could be done

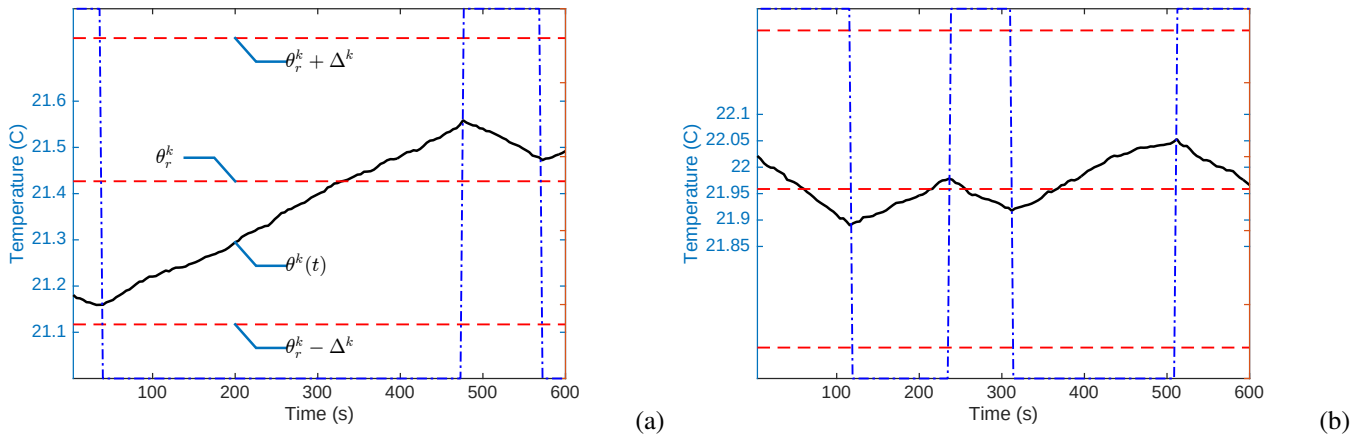


Fig. 4: Temperature profile of 2 individual TCLs under priority-based stack control with no-short-cycling constraint. Units have different temperature set-points and boundaries as well as quite different ON/OFF cycling patterns (the blue dot-dashed line).

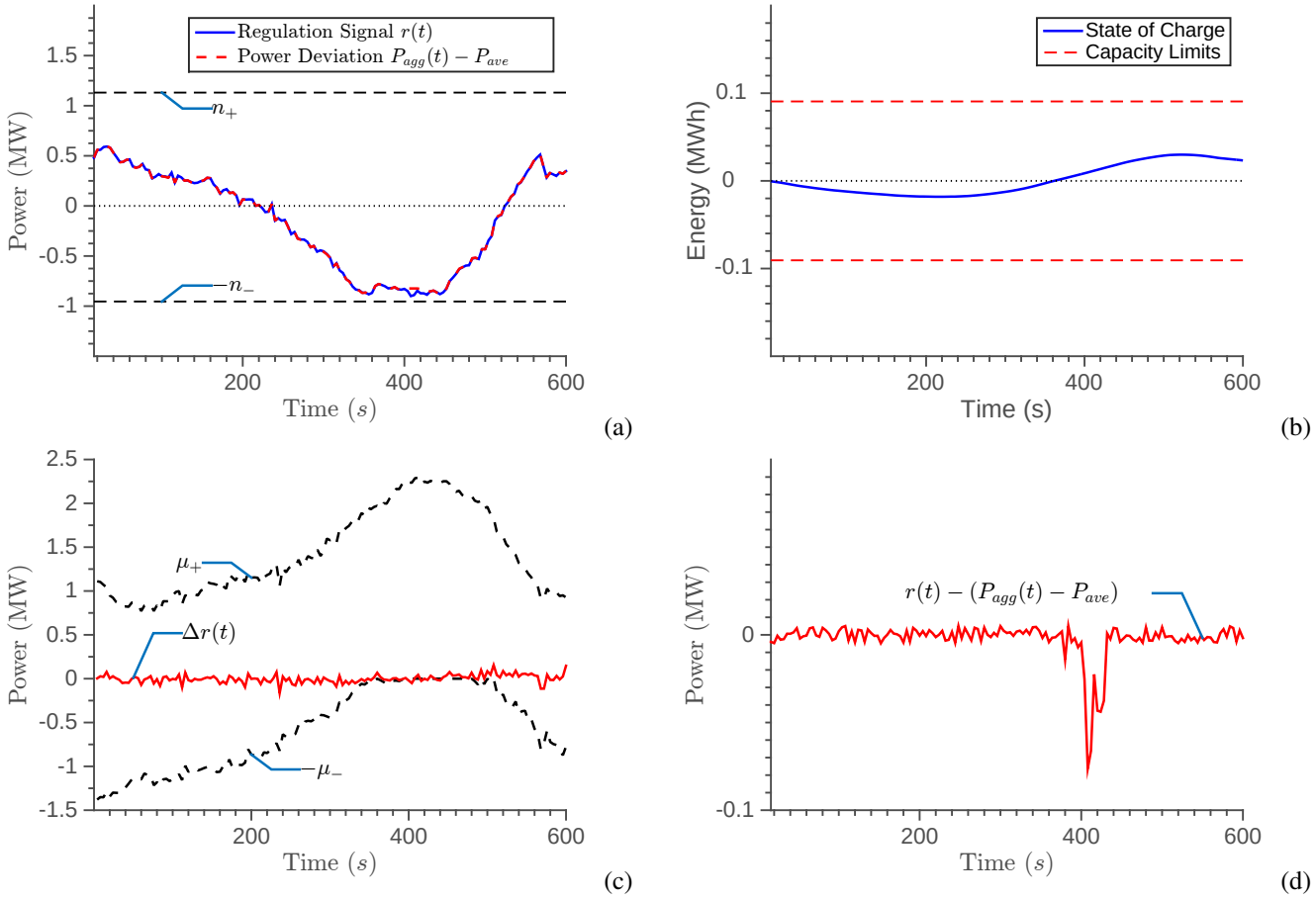


Fig. 5: Imposing the no-short-cycling requirement on TCLs and its impact on ramping rate capability of the collection in providing regulation. (a) The regulation signal and battery model bounds on power. (b) The SoC and the capacity limits. (c) $\Delta r(t)$ and its bounds given in Theorem 2. (d) The difference between the desired regulation signal $r(t)$ and the actual power draw $P_{agg}(t) - P_{ave}$. Around time 400 (s), $-\mu_-(t)$ approaches zero, meaning that negative $\Delta r(t)$ is no longer feasible. Fig. 5(d) confirms that the regulation signal is not well followed downward during the time that $-\mu_-(t)$ is close to zero.

in much the same manner that AGC is provided. A normalized "ramping signal" is transmitted by the system operator, and resources must follow this signal proportional to the ramping capacity they commit to in a forward ancillary services market.

APPENDIX A

In this appendix, we briefly summarize the GBM. This battery model is a core component in characterizing the aggregate flexibility of a collection of TCLs. An aggregate model cannot be created by simply summing the power limits and capacities of individual GBMs. In the following, we provide counterexamples. Recall that each TCL can accept perturbations around its nominal power consumption ($p^k(t) = P_o^k + e^k(t)$) that will meet user-specified comfort bounds. Define

$$\mathbb{E}^k := \left\{ e^k(t) \Big| \begin{matrix} 0 \le P_o^k + e^k(t) \le P_m^k, \\ P_o^k + e^k(t) \text{ maintains } |\theta^k(t) - \theta_r^k| \le \Delta^k \end{matrix} \right\}.$$

This set of power signals represents the flexibility of the kth TCL with respect to its nominal. The *aggregate flexibility* of the collection of TCLs is then defined as the Minkowski sum

$$\mathbb{U} = \sum_k \mathbb{E}^k.$$

It is hard to directly evaluate set $\mathbb{U}$. It has been shown that $\mathbb{U}$ can be nested between two other sets using GBMs [37].

Definition 3 ([37]): Let $\phi = (n_-, n_+, \mathcal{C})$ be non-negative parameters. For a given $\alpha > 0$, a Generalized Battery Model (GBM) $\mathbb{B}(\phi)$ is a set of signals $u(t)$ that satisfy

$$u(t) \in \mathbb{B}(\phi) \Longleftrightarrow \begin{cases} -n_- \le u(t) \le n_+ \\ \dot{x}(t) = -\alpha x(t) - u(t) \\ x(0) = 0 \implies |x(t)| \le \mathcal{C},\ \forall t > 0. \end{cases} \quad \square$$

Each set of parameters $(n_-, n_+, \mathcal{C})$ and α define a new GBM. Let's consider the following 3 GBMs.

$$S^1 = \left\{ u^1(t) \left| \begin{matrix} -n_-^1 \le u^1(t) \le n_+^1 \\ \dot{x}^1 = -\alpha^1 x^1 - u^1,\ x^1(0) = 0 \\ \implies |x^1(t)| < \mathcal{C}^1 \end{matrix} \right. \right\},$$

$$S^2 = \left\{ u^2(t) \left| \begin{matrix} -n_-^2 \le u^2(t) \le n_+^2 \\ \dot{x}^2 = -\alpha^2 x^2 - u^2,\ x^2(0) = 0 \\ \implies |x^2(t)| < \mathcal{C}^2 \end{matrix} \right. \right\},$$

$$S^3 = \left\{ u^3(t) \left| \begin{matrix} -n_-^1 - n_-^2 \le u^3(t) \le n_+^1 + n_+^2 \\ \dot{x}^3 = -\alpha^3 x^3 - u^3, x^3(0) = 0 \\ \implies |x^3(t)| < \mathcal{C}^1 + \mathcal{C}^2 \end{matrix} \right. \right\}.$$

While the parameters of S^3 are sum of the parameters of S^1 and S^2, in the following we show that $S_1 + S_2 \neq S_3$. In order to do that, we first give a counterexample showing that $S_1 + S_2 \not\subseteq S_3$. Choose $\alpha^1 = \alpha^3 < \alpha^2$ and $\mathcal{C}^i \alpha^i < n_+^i$ for $i = 1, 2$. Let $u^1(t) = \alpha^1 \mathcal{C}^1$ and $u^2(t) = \alpha^2 \mathcal{C}^2$ (constant signals). Note that the steady state response to these signals is $\lim_{t\to\infty} x^1(t) = \mathcal{C}^1$ and $\lim_{t\to\infty} x^2(t) = \mathcal{C}^2$, implying that $u^1(t) \in S^1$ and $u^2(t) \in S^2$. However, with $u^3(t) = u^1(t) + u^2(t)$, we get $\lim_{t\to\infty} x^3(t) = \frac{\alpha^1 \mathcal{C}^1 + \alpha^2 \mathcal{C}^2}{\alpha^3} > \mathcal{C}^1 + \mathcal{C}^2$, thus $u^3(t) \notin S_3$. Therefore, $S_1 + S_2 \not\subseteq S_3$.

We then show that $S_3 \not\subseteq S_1 + S_2$. Let's consider lossless batteries (i.e., $\alpha_1 = \alpha_2 = \alpha_3 = 0$) with different capacities and power limits. One battery has a large capacity but small power limit, while the other has a small capacity and large power limit. These two batteries cannot be modeled by a battery with the sum of the capacities and sum of the power limits. In particular, let's choose $n_+^1 > n_+^2$ and $\mathcal{C}^1 < \mathcal{C}^2$. Let $A = n_+^1 + n_+^2$ and $\tau = (\mathcal{C}^1 + \mathcal{C}^2)/(n_+^1 + n_+^2)$. Consider

$$u^3(t) = \begin{cases} A, & 0 \le t \le \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $u^3(t) \in S^3$. We now look for $u^1(t) \in S^1$ and $u^2(t) \in S^2$ such that $u^1(t) + u^2(t) = u^3(t)$. We will check this at time $t = \tau$. Let $u^2(t) = A - u^1(t)$, and check if there exists $u^1(t)$ that satisfies the following inequalities:

$$-n_-^1 \le u^1(t) \le n_+^1, \tag{6a}$$

$$-\mathcal{C}^1 \le \int_0^\tau u^1(t)dt \le \mathcal{C}^1, \tag{6b}$$

$$-n_-^2 \le A - u^1(t) \le n_+^2, \tag{6c}$$

$$-\mathcal{C}^2 \le \int_0^\tau (A - u^1(t))dt \le \mathcal{C}^2. \tag{6d}$$

Since $A = n_+^1 + n_+^2$, (6c) implies

$$u^1(t) \ge n_+^1,$$

which along with (6a) implies $u^1(t) = n_+^1$. Then,

$$\int_0^\tau u^1(t)dt = \tau n_+^1 = n_+^1(\mathcal{C}^1 + \mathcal{C}^2)/(n_+^1 + n_+^2).$$

Since $n_+^1 > n_+^2$ and $\mathcal{C}^1 < \mathcal{C}^2$

$$\int_0^\tau u^1(t)dt = n_+^1(\mathcal{C}^1 + \mathcal{C}^2)/(n_+^1 + n_+^2) > \frac{1}{2}(\mathcal{C}^1 + \mathcal{C}^2) > \mathcal{C}^1,$$

which is incompatible with (6b). $\square$

APPENDIX B

Proof of Theorem 1

1) Applying $p(t) = \frac{1}{N}\sum_k q^k(t)$ to the continuous-power model and using linearity, it is trivial to show that $\theta_c(t) = \frac{1}{N}\sum_k \theta^k(t)$. $\theta^k(t) \in [\theta_r - \Delta, \theta_r + \Delta], \forall k$. Thus, $\theta_c(t) = \frac{1}{N}\sum_k \theta^k(t) \in [\theta_r - \Delta, \theta_r + \Delta]$ and this completes the proof.

2) The proof is by induction. Let $\theta_{agg}(t) = \frac{1}{N}\sum_k \theta^k(t)$. Clearly $\|\theta_c(0) - \theta_{agg}(0)\| \le \frac{\epsilon}{N}$. Now, let n be an arbitrary integer and let $t_0 = nT$. Suppose $\|\theta_c(t_0) - \theta_{agg}(t_0)\| \le \frac{\epsilon}{N}$. For $t \in (t_0, t_0 + T)$, without loss of generality, suppose $r(t) > 0$, and also assume that units are ordered based on their temperature from the highest to the lowest such that $\theta^1(t) > \theta^2(t) > \cdots > \theta^N(t)$. Set $q^k(t) = 1$ until (i) $|\frac{1}{N}\sum_k q^k(t) - p(t)| \le \frac{1}{N}$ and (ii) $\text{sign}(\frac{1}{N}\sum_k q^k(t) - p(t)) = \text{sign}(\theta_{agg}(t) - \theta_c(t))$. By linearity,

$$\dot{\theta}_{agg}(t) = -a(\theta_{agg}(t) - \theta_a) - bP_m \frac{1}{N}\sum_k q^k(t). \tag{7}$$

Based on (3) and (7), we have

$$\theta_{agg}(t_0 + T) - \theta_c(t_0 + T) =$$

$$e^{-aT}(\theta_{agg}(t_0)-\theta_c(t_0))-(1-e^{-aT})(\frac{bP_m}{a})(\frac{\sum_k q^k(t)}{N}-p(t)).$$

Since $\left|(1-e^{-aT})(\frac{bP_m}{a})(\frac{\sum_k q^k(t)}{N}-p(t))\right| \leq \epsilon/N$ and the fact that $\text{sign}(\frac{1}{N}\sum_k q^k(t)-p(t)) = \text{sign}(\theta_{agg}(t)-\theta_c(t))$, we have

$$\left|(\theta_{agg}(t_0+T)-\theta_c(t_0+T))\right| \leq \frac{\epsilon}{N}.$$

Since n was arbitrary,

$$|\theta_{agg}(t)-\theta_c(t)| \leq \frac{\epsilon}{N} \tag{8}$$

for all $t > 0$ by induction. Now, let $\theta_{\max}(t)$ and $\theta_{\min}(t)$ denote the temperature of units with maximum and minimum temperature at time t, respectively. Let $k_{\max}$ and $k_{\min}$ denote the units with maximum and minimum temperature at time $t+T$, respectively. Then based on the deadband model (1)

$$\theta_{\max}(t+T)-\theta_{\min}(t+T) = \theta^{k_{\max}}(t+T)-\theta^{k_{\min}}(t+T) =$$
$$e^{-aT}(\theta^{k_{\max}}(t)-\theta^{k_{\min}}(t))-(1-e^{-aT})\frac{P_m b}{a}(q^{k_{\max}}(t)-q^{k_{\min}}(t)).$$

Since units were turned ON in temperature order, $q^{k_{\max}}(t) - q^{k_{\min}}(t)$ is either zero or it has magnitude 1 with the same sign as $\theta^{k_{\max}}(t)-\theta^{k_{\min}}(t)$. Note that in this case only the unit with a higher temperature is ON. Thus,

$$|\theta_{\max}(t+T)-\theta_{\min}(t+T)| \leq$$
$$\max\left(e^{-aT}|\theta^{k_{\max}}(t)-\theta^{k_{\min}}(t)|, (1-e^{-aT})\frac{P_m b}{a}\right)$$
$$\leq \max(e^{-aT}|\theta_{\max}(t)-\theta_{\min}(t)|, \epsilon).$$

Since this is true for all T, and $|\theta_{\max}(0)-\theta_{\min}(0)| \leq 2\Delta$,

$$|\theta_{\max}(t+T)-\theta_{\min}(t+T)| \leq \max\left(2e^{-aT}\Delta, \epsilon\right). \tag{9}$$

Bounds (8) and (9), and the fact that $\theta_i(t)-\theta_{\text{agg}}(t) < \theta_{\max}(t)-\theta_{\min}(t)$ for all time gives the result.

APPENDIX C

Proof of Theorem 2 At a given time t, let $P_{\text{ON}}(t)$ and $P_{\text{OFF}}(t)$ denote the total power of ON and OFF units, respectively. Let $r(t)$ be the AGC signal. If $r(t)$ is satisfied, then by definition

$$P_{\text{ON}}(t) = P_{\text{ave}} + r(t),$$
$$P_{\text{OFF}}(t) = P_{\text{tot}} - P_{\text{ave}} - r(t).$$

Note that the second equality follows from the fact that $P_{\text{ON}}(t) + P_{\text{OFF}}(t) = \sum_k P_m^k := P_{\text{tot}}$. At a given time t, let $P_{\text{ON}}^{\text{unavail}}(t)$ and $P_{\text{OFF}}^{\text{unavail}}(t)$ be the total power of unavailable ON and OFF units, respectively. Clearly,

$$P_{\text{ON}}^{\text{avail}}(t) = P_{\text{ON}}(t) - P_{\text{ON}}^{\text{unavail}}(t),$$
$$P_{\text{OFF}}^{\text{avail}}(t) = P_{\text{OFF}}(t) - P_{\text{OFF}}^{\text{unavail}}(t).$$

Also, $P_{\text{ON}}^{\text{unavail}}$ is given by the sum of the power of units that have been turned ON in the last τ sample times. Formally,

$$P_{\text{ON}}^{\text{unavail}}(t) = \sum_{k=t-\tau}^{t} P_{\text{OFF}\to\text{ON}}(k), \tag{10}$$

where $P_{\text{OFF}\to\text{ON}}(k)$ is the power of units that turn ON at time k. Part of such units turn ON because of the local deadband controller (denoted by $P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k)$). More OFF to ON power should be considered based on the required regulation change at time k (i.e., $\Delta r(k)$) and the potential imbalance between $P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k)$ and $P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k)$.

$$D(k) := \Delta r(k) - (P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) - P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k)).$$

Observe that when $\Delta r(k) = P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) - P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k)$, then no manipulation of TCLs is required and therefore,

$$P_{\text{OFF}\to\text{ON}}(k) = P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k),$$
$$P_{\text{ON}\to\text{OFF}}(k) = P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k).$$

If $\Delta r(k) > \left(P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) + (-P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k))\right)$, then we need extra power compensation from the collection. Therefore,

$$\begin{aligned} P_{\text{OFF}\to\text{ON}}(k) &= P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) + D(k) \\ &= \Delta r(k) + P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k) \\ &= P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) + [D(k)]_+ . \end{aligned} \tag{11}$$

Similarly, $P_{\text{OFF}}^{\text{unavail}}$ is the sum of the power of units that have been turned OFF in the last τ sample times. Formally,

$$P_{\text{OFF}}^{\text{unavail}}(t) = \sum_{k=t-\tau}^{t} P_{\text{ON}\to\text{OFF}}(k), \tag{12}$$

where $P_{\text{ON}\to\text{OFF}}(k)$ is the power of units that turn OFF at time k. Part of such units turn OFF because of the local deadband controller (denoted by $P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k)$). More ON to OFF manipulation is needed based on the required regulation change at time k (i.e., $\Delta r(k)$) and the potential imbalance between $P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k)$ and $P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k)$. If $\Delta r(k) < \left(P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) + (-P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k))\right)$, then $\left(P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) + (-P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k)) - \Delta r(k)\right)$ less power is required from the collection. As a result,

$$\begin{aligned} P_{\text{ON}\to\text{OFF}}(k) &= P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k) + (-D(k)) \\ &= P_{\text{OFF}\to\text{ON}}^{\text{lim}}(k) - \Delta r(k) \\ &= P_{\text{ON}\to\text{OFF}}^{\text{lim}}(k) + [-D(k)]_+ . \end{aligned}$$

Note this derivation is agnostic to the sign of $\Delta r(k)$ and is valid for either $\Delta r(k) > 0$ or $\Delta r(k) < 0$. Also, we can consider different rest time for ON and OFF states. The results follow immediately just by substituting τ with appropriate rest time for each scenario.

APPENDIX D

Proof of Theorem 3 Let's first consider the OFF units. As mentioned earlier, the total power of OFF units at time $t-1$ is given by $P_{\text{OFF}}(t-1) = P_{\text{OFF}}^{\text{avail}}(t-1) + P_{\text{OFF}}^{\text{unavail}}(t-1)$. If we assume none of the unavailable OFF units at time $t-1$ become available at time t, then $P_{\text{OFF}}^{\text{avail}}(t) = P_{\text{OFF}}^{\text{avail}}(t-1)$. Also, let's assume all units that just hit the temperature boundaries (either the upper or the lower limit) are unavailable at time t. Under these two assumptions, an upper bound on $r(t)-r(t-1)$ can be achieved under the worst case scenario assumptions as

$$r(t)-r(t-1) \leq P_{\text{OFF}}^{\text{avail}}(t-1) - \max(P_{\text{ON}\to\text{OFF}}^{\text{lim}}(t), P_{\text{OFF}\to\text{ON}}^{\text{lim}}(t)).$$

Similarly, if all units that just hit the temperature limits are currently unavailable and none of the unavailable ON units at time $t-1$ become available at time t, an upper bound on $r(t-1)-r(t)$ can be achieved under the worst case scenario assumptions as

$$r(t-1)-r(t) \leq P_{\text{ON}}^{\text{avail}}(t-1)-\max(P_{\text{ON}\rightarrow\text{OFF}}^{\text{lim}}(t), P_{\text{OFF}\rightarrow\text{ON}}^{\text{lim}}(t)).$$

ACKNOWLEDGMENT

The authors gratefully acknowledge He Hao, Pravin Varaiya, and all anonymous reviewers who have contributed to this paper with their insightful comments and discussions.

REFERENCES

[1] Office of Energy Policy and Innovation, "Integration of variable energy resources - FERC Order No. 764," Federal Energy Regulatory Commission – FERC, Washington, D.C., Tech. Rep. Docket No. RM10-11-000, 2012. [Online]. Available: http://www.ferc.gov/whats-new/comm-meet/2012/062112/E-3.pdf

[2] B. Kirby, "Ancillary services: Technical and commercial insights," Wärtsilä North America, Inc., Huston, TX, Tech. Rep., July 2007.

[3] B. M. Sanandaji, H. Hao, and K. Poolla, "Fast regulation service provision via aggregation of thermostatically controlled loads," *in Proceedings of the 47th Hawaii International Conference on System Sciences – HICSS47*, pp. 2388–2397, 2014.

[4] J. C. Smith, M. R. Milligan, E. A. DeMeo, and B. Parsons, "Utility wind integration and operating impact state of the art," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 900–908, 2007.

[5] Y. V. Makarov, C. Loutan, J. Ma, and P. de Mello, "Operational impacts of wind generation on California power systems," *IEEE Transactions on Power Systems*, vol. 24, no. 2, pp. 1039 –1050, 2009.

[6] S. Meyn, M. Negrete-Pincetic, G. Wang, A. Kowli, and E. Shafieepoorfard, "The value of volatile resources in electricity markets," *in Proceedings of the 49th IEEE Conference on Decision and Control – CDC*, pp. 1029–1036, 2010.

[7] U. Helman, "Resource and transmission planning to achieve a 33% RPS in California–ISO modeling tools and planning framework," in *FERC Technical Conference on Planning Models and Software (Vol. 2010)*, 2010.

[8] Market and Infrastructure Policy, "2013 flexible capacity procurement requirement," California Independent System Operator–CAISO, Folsom, CA, Tech. Rep., March 2012. [Online]. Available: https://www.caiso.com/Documents/2013FlexibleCapacityProcurementRequirementProposalSupplement.pdf

[9] K. Vu, R. Masiello, and R. Fioravanti, "Benefits of fast-response storage devices for system regulation in ISO markets," *IEEE Power Energy Society General Meeting*, pp. 1–8, 2009.

[10] Y. V. Makarov, L. S., J. Ma, and T. B. Nguyen, "Assessing the value of regulation resources based on their time response characteristics," Pacific Northwest National Laboratory, Richland, WA, Tech. Rep. PNNL-17632, 2008.

[11] Office of Energy Policy and Innovation, "Frequency regulation compensation in the organized wholesale power markets - FERC Order No. 755," Federal Energy Regulatory Commission – FERC, Washington, D.C., Tech. Rep. Docket Nos. RM11-7-000 and AD10-11-000, 2011. [Online]. Available: https://www.ferc.gov/whats-new/comm-meet/2011/102011/E-28.pdf

[12] ——, "Third-party provision of ancillary services; accounting reporting for new electric storage technologies - FERC Order No. 784," Federal Energy Regulatory Commission – FERC, Washington, D.C., Tech. Rep. Docket Nos. RM11-24-000 and AD10-13-000, 2013. [Online]. Available: https://www.ferc.gov/whats-new/comm-meet/2013/071813/E-22.pdf

[13] California Independent System Operator, "Demand response and energy efficiency roadmap: Maximizing preferred resources," CAISO, Folsom, CA, Tech. Rep., December 2013. [Online]. Available: http://www.caiso.com/Documents/DR-EERoadmap.pdf

[14] D. S. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy," *Energy Conversion and Management*, vol. 50, no. 5, pp. 1389–1400, 2009.

[15] S. Koch, M. Zima, and G. Andersson, "Active coordination of thermal household appliances for load management purposes," *Proceedings of the IFAC Symposium on Power Plants and Power Systems Control*, pp. 149–154, 2009.

[16] S. Koch, J. Mathieu, and D. Callaway, "Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services," *Proceedings of the 17th Power Systems Computation Conference – PSCC*, pp. 1–7, 2011.

[17] J. Mathieu and D. Callaway, "State estimation and control of heterogeneous thermostatically controlled loads for load following," *Proceedings of the 45th Hawaii International Conference on System Sciences – HICCS45*, pp. 2002–2011, 2012.

[18] J. L. Mathieu, S. Koch, and D. S. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 430 –440, February 2013.

[19] J. L. Mathieu, M. Kamgarpour, J. Lygeros, and D. S. Callaway, "Energy arbitrage with thermostatically controlled loads," in *European Control Conference (ECC)*, 2013, pp. 2519–2526.

[20] G. Koutitas, "Control of flexible smart devices in the smart grid," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1333–1343, 2012.

[21] W. Zhang, K. Kalsi, J. Fuller, M. Elizondo, , and D. Chassin, "Aggregate model for heterogeneous thermostatically controlled loads with demand response," *Proceedings of the 2012 IEEE PES General Meeting*, pp. 1–8, 2012.

[22] C.-Y. Chang, W. Zhang, J. Lian, and K. Kalsi, "Modeling and control of aggregated air conditioning loads under realistic conditions," *Proceedings of the IEEE PES Innovative Smart Grid Technologies Conference – ISGT*, pp. 1–6, 2013.

[23] Y. Zhang and N. Lu, "Parameter selection for a centralized thermostatically controlled appliances load controller used for intra-hour load balancing," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 2100–2108, 2013.

[24] W. Zhang, J. Lian, C.-Y. Chang, and K. Kalsi, "Aggregated modeling and control of air conditioning loads for demand response," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4655 – 4664, 2013.

[25] N. Lu, "An evaluation of the HVAC load potential for providing load balancing service," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1263–1270, 2012.

[26] S. Bashash and H. K. Fathy, "Modeling and control insights into demand-side energy management through setpoint control of thermostatic loads," *Proceedings of the 2011 American Control Conference – ACC*, pp. 4546–4553, 2011.

[27] ——, "Modeling and control of aggregate air conditioning loads for robust renewable power management," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 4, pp. 1318–1327, 2013.

[28] Y. Lin, P. Barooah, S. Meyn, and T. Middelkoop, "Experimental evaluation of frequency regulation from commercial building hvac systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 776–783, 2015.

[29] S. Bashash and H. K. Fathy, "Robust demand-side plug-in electric vehicle load control for renewable energy management," *Proceedings of the 2011 American Control Conference – ACC*, pp. 929–934, 2011.

[30] A. Nayyar, J. Taylor, A. Subramanian, K. Poolla, and P. Varaiya, "Aggregate flexibility of a collection loads," *in Proceedings of the 52th IEEE Conference on Decision and Control – CDC*, pp. 5600–5607, 2013.

[31] M. Maasoumy, B. M. Sanandaji, K. Poolla, and A. Sangiovanni-Vincentelli, "Model predictive control of regulation services from commercial buildings to the smart grid," *in Proceedings of the 2014 American Control Conference – ACC*, pp. 2226–2233, 2014.

[32] J. Qin, Y. Chow, J. Yang, and R. Rajagopal, "Modeling and online control of generalized energy storage networks," *in Proceedings of the 5th international conference on future energy systems*, pp. 27–38, 2014.

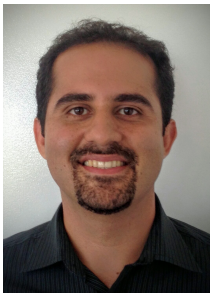
[33] M. Alizadeh, A. Scaglione, A. Applebaum, G. Kesidis, and K. Levitt, "Scalable and anonymous modeling of large populations of flexible appliances," *arXiv preprint arXiv:1404.1958*, 2014.

[34] S. Kundu, N. Sinitsyn, S. Backhaus, and I. Hiskens, "Modeling and control of thermostatically controlled loads," *Proceedings of the 17th Power Systems Computation Conference – PSCC*, pp. 1–7, 2011.

[35] Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE). (2011) Buildings energy data book. [Online]. Available: http://buildingsdatabook.eren.doe.gov/default.aspx

[36] "U.S. Energy Information Administration (EIA), annual energy review," 2010. [Online]. Available: http://www.eia.gov/totalenergy/data/annual/#consumption

[37] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Aggregate flexibility of thermostatically controlled loads," *IEEE Transactions of Power Systems*, vol. 30, pp. 189–198, January 2015.
[38] ——, "A Generalized Battery Model of a Collection of Thermostatically Controlled Loads for Providing Ancillary Service," *in Proceedings of the 51th Annual Allerton Conference on Communication, Control and Computing*, pp. 551–558, 2013.
[39] R. Malhame and C.-Y. Chong, "Electric load model synthesis by diffusion approximation of a high-order hybrid-state stochastic system," *IEEE Transactions on Automatic Control*, vol. 30, no. 9, pp. 854–860, 1985.
[40] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Frequency regulation from flexible loads: Potential, economics, and implementation," *in Proceedings of the 2014 American Control Conference – ACC*, pp. 65–72, 2014.
[41] ——, "Potentials and economics of residential thermal loads providing regulation reserve," *Energy Policy*, vol. 79, pp. 115–126, January 2015.
[42] N. Lu and Y. Zhang, "Design considerations of a centralized load controller using thermostatically controlled appliances for continuous regulation reserves," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 914–921, 2013.
[43] P. Barooah, A. Buic, and S. Meyn, "Spectral decomposition of demand-side flexibility for reliable ancillary services in a smart grid," *48th Hawaii International Conference on System Sciences – HICSS*, pp. 2700–2709, 2015.
[44] "PJM Regulation Data," PJM Interconnection, Audubon, PA. [Online]. Available: http://www.pjm.com/markets-and-operations/ancillary-services/mkt-based-regulation/fast-response-regulation-signal.aspx



Borhan M. Sanandaji is a postdoctoral scholar at the University or California, Berkeley in the Electrical Engineering and Computer Sciences department. He received his Ph.D. degree (2012) from the Colorado School of Mines and his B.Sc. degree (2004) from the Amirkabir University of Technology (Tehran, Iran), all in electrical engineering. His current research interests include big physical data analytics, compressive sensing and structured-sparse recovery, and low-dimensional modeling with applications in energy systems, renewable energy integration and demand response, control theory, cyber security and fault detection, and intelligent transportation systems.

Tyrone L. Vincent received the B.S. degree in electrical engineering from the University of Arizona, Tucson, in 1992, and the M.S. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, in 1994 and 1997, respectively. He is currently a Professor in the Department of Electrical Engineering and Computer Science, at the Colorado School of Mines, Golden. His research interests include system identification, estimation and control with applications in materials processing and energy systems.

Kameshwar Poolla received his BTech from the Indian Institute of Technology, Bombay in 1980 and his PhD from the University of Florida in 1984. He is currently the Cadence Distinguished Professor at UC Berkeley in EECS and ME. His current research interests include many aspects of future energy systems including economics, security, and commercialization. He also serves as the Founding Director of the IMPACT Center for Integrated Circuit manufacturing. Dr. Poolla co-founded OnWafer Technologies which was acquired by KLA-Tencor in 2007. Dr. Poolla has been awarded a 1988 NSF Presidential Young Investigator Award, the 1993 Hugo Schuck Best Paper Prize, the 1994 Donald P. Eckman Award, the 1998 Distinguished Teaching Award of the University of California, the 2005 and 2007 IEEE Transactions on Semiconductor Manufacturing Best Paper Prizes, and the 2009 IEEE CSS Transition to Practice Award.