Aggregate Flexibility of Thermostatically Controlled Loads[π]

He Hao[a], Borhan M. Sanandaji[b], Kameshwar Poolla[c], and Tyrone L. Vincent[d]

***Abstract*— It is widely accepted that Thermostatically Controlled Loads (TCLs) can be used to provide regulation reserve to the grid. We first argue that the aggregate flexibility offered by a collection of TCLs can be succinctly modeled as a *stochastic battery* with dissipation. We next characterize the power limits and energy capacity of this battery model in terms of TCL parameters and random exogenous variables such as ambient temperature and user-specified set-points. We then describe a direct load control architecture for regulation service provision. Here, we use a priority-stack-based control framework to select which TCLs to control at any time. The control objective is for the aggregate power deviation from baseline to track an automatic generation control signal supplied by the system operator. Simulation studies suggest the practical promise of our methods.**

I. Introduction

Ancillary services provide for resources to handle supply-demand imbalances at various time-scales, maintain power quality, and assure reliable power delivery under contingencies. There are many different ancillary services. Of these, regulation reserve (or frequency regulation) and load-following are the key services for maintaining the power balance under normal operating conditions. Load following is a feed-forward approach and handles predictable and slower changes in load. Regulation is a feedback strategy that mitigates faster and unpredictable changes in system load and corrects unintended fluctuations in generation [1].

It is widely accepted that the deep penetration of renewable generation will substantially increase the need for ancillary services [2]–[4]. Recent studies [5] project that the spring time maximum up-regulation reserve needed to accommodate California's 33% renewable penetration mandate will increase from 277 MW to 1,135 MW. Similar increases in down-regulation reserve are projected. The maximum load-following requirement will need to increase from 2.3 GW to 4.4 GW. If these additional ancillary services are provided by traditional fossil fuel generators, it will diminish the net carbon benefit from renewables, reduce generation efficiency, and be economically untenable.

[a]Electrical Engineering & Computer Sciences, UC Berkeley, CA 94720. Corresponding author. hehao@berkeley.edu.
[b]Electrical Engineering & Computer Sciences, UC Berkeley, CA 94720.
[c]Electrical Engineering & Computer Sciences, UC Berkeley, CA 94720.
[d]Electrical Engineering & Computer Science, Colorado School of Mines, Golden, CO 80401.
[π]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.

There is an emerging consensus that demand side resources must play a key role in supplying zero-emission regulation service that is necessary for deep renewable integration. These include thermostatically controlled loads, electric vehicles, and strategic storage. Existing programs, such as the SmartAC™ program of PG&E, aggregate residential air conditioners for peak load shaving and emergency load management [6]. Because these load control mechanisms are primarily concerned with very low frequency changes in demand (i.e., the changes occur over hours timescale), they are invoked infrequently and offer limited financial value. In contrast, there is an enormous untapped potential for flexible loads to offer more lucrative fast ancillary services such as frequency regulation or load-following.

Residential Thermostatically Controlled Loads (TCLs) such as air conditioners, heat pumps, water heaters, and refrigerators, represent about 20% of the total electricity consumption in the United States [7], [8], and thus offer significant potential for provision of various ancillary services. TCLs have inherent thermal storage, so their electricity consumption can be modulated while still meeting the desired temperature requirements of the end user.

This paper aims to provide a foundation for a practical method by which TCLs can be utilized to provide regulation reserve to the grid. In the scheme we consider, an aggregator would contract with owners of individual TCLs to *directly control* the electricity consumption of their units. Units owners are assured that the temperature limits they specify through thermostat set-point selection will be respected. We provide an important tool that would be useful to the aggregator: *a simple, compact battery model* that characterizes the set of power profiles that the collection of TCLs can accept while meeting their local constraints. This model depends on exogenous random processes such as ambient temperatures, user specified set-points, and unit participation. The aggregator could forecast these random processes and predict the admissible set of power profiles for the TCL collection on a future window. This flexibility in power consumption is monetized by offering it in ancillary service markets to provide regulation reserve. If the offer is accepted, the aggregator is obligated to deliver the service at delivery time. The aggregator then uses a *priority stack control strategy* to track an exogenous regulation signal supplied by the system operator.

A. Main Contributions

Thermostatically controlled loads are flexible in the sense that a variety of power trajectories are able to meet user-

specified temperature constraints. We consider a heterogeneous collection of TCLs indexed by k. We describe TCLs with a simplified continuous power model. For each TCL, we define a nominal power profile P_o^k and model its flexibility as the set $\mathbb{E}^k$ of permissible deviations from this nominal that result in temperature trajectories which respect the dead-band constraints of the unit. The aggregate flexibility $\mathbb{U}$ of the collection is the sum of the flexibilities of the individual units, i.e. $\mathbb{U} = \sum_k \mathbb{E}^k$. This set has a complex structure.

We first introduce a generalized battery model $\mathbb{B}(\phi)$ which serves as a compact, abstract representation of a set of power signals. The battery model $\mathbb{B}(\phi)$ has parameters $\phi = (C, n_-, n_+, \alpha)$, where C is the battery energy capacity, n_- and n_+ are the charge/discharge rate limits, and α is the dissipation rate. This simple model is attractive as it is defined by a few physically intuitive parameters. The main contribution of this paper is to establish that the set $\mathbb{U}$ can be bounded by two generalized battery models:

$$\mathbb{B}(\phi_1) \subseteq \mathbb{U} \subseteq \mathbb{B}(\phi_2).$$

We offer explicit formulae for the battery parameters ϕ_1 and ϕ_2 in terms of individual TCL parameters and exogenous random processes such as ambient temperature and user defined set-points. We further show that the gap between these two models vanishes as the population of TCLs becomes homogeneous.

We next offer a direct-load control architecture for an aggregator to provide frequency regulation service from a collection of TCLs. This requires an *ex ante* contractual agreement with the system operator that specifies two items on the forward delivery window: (a) the "nominal" or baseline power consumption $n(t)$ of the aggregate, and (b) the class of admissible regulation signals $r(t)$. In the event the offered service is accepted, the aggregator receives an AGC signal from the system operator and is obligated to ensure that the collection of TCLs consumes power profile $r(t) + n(t)$ in aggregate. The aggregator is also obligated to ensure that power profiles allocated to each TCL must respect their temperature dead-band constraints. If $r(t) \in \mathbb{B}(\phi_1)$, this is possible. If $r(t) \notin \mathbb{B}(\phi_2)$, this is not possible. Thus, the aggregator will forecast the battery parameters ϕ_1 and offer the regulation capability $\mathbb{B}(\phi_1)$ which it is assured that it can reliably deliver.

If the offered service is accepted, the aggregator receives an AGC signal $r(t) \in \mathbb{B}(\phi_1)$ from the system operator, and must devise a *run-time* strategy to allocate the power $r(t) + n(t)$ to the various TCLs. This allocation must be *causal* and meet the user-defined temperature needs of the TCLs. We propose an allocation strategy based on priority stacks. Priorities are computed based on temperature distance to switching boundary or time to reach the switching boundary. The priority-stack-based strategy naturally accommodates operational constraints such as avoiding short-cycling on each unit, and most importantly offers a generic architecture suitable for direct load control of various flexible loads including electric vehicles and deferrable appliances. We explore some details of the practical implementations of our control strategy, and illustrate our principal ideas with simulation studies.

Our *set-theoretic characterization* of the aggregate flexibility of a collection of TCLs using battery models allows us to naturally treat regulation service provision. Ex ante, it is used to conservatively represent the set of feasible regulation signals that the aggregator can support. The resulting run-time control problem to follow the external regulation signal r becomes *extremely simple*. There is no need to account for the dynamics of the TCL population.

A preliminary version of this paper was presented in [9].

B. Related Work

Early work showing potential of TCLs for mitigating the variability from renewables may be found in [10]. Here, the author shows that the aggregate power consumption of a collection of TCLs can be made to follow a high-frequency signal such as the power output of a wind-farm.

There is a substantial body of work on models for TCLs and TCL aggregations [11]–[16]. These models fall along a spectrum of complexity and fidelity, and have been used for a variety of simulations, control design, and validation efforts. There are some papers that propose battery models to treat the aggregate flexibility for a collection of TCLs, see for example [17]. These papers do not qualify their battery models as approximations, nor do they attempt to quantify the quality of the approximation. The battery models derived usually do not include a dissipation term and therefore understate the aggregate flexibility of TCL collections. To the best of our knowledge, our set-theoretic characterization of aggregate flexibility together with quantifying the associated error using generalized battery models is new. Temperature-based and/or time-based priority control methods that are closely related or identical to this paper were developed in [15], [18], [19].

Another class of models commonly used to study aggregations of TCLs are population-bin transition models, for example [15], [17]. These control-oriented models are often high order (depending on the number of bins), linear ODEs. These researchers then study regulation as a control system design problem with this model. There are three difficulties with population bin models: (a) they are complex, not universal portable, (b) they do not easily handle participation effects (i.e. modifying the model to account for individual TCLs coming on line or dropping out), (c) they require much more complex control strategies that take into account population dynamics.

While there is considerable literature on *indirect load control* through price proxies [20], [21], it is our view that these schemes are less reliable, as resource availability and tracking accuracy of a dispatch signal are not guaranteed [22]. Moreover, with large scale deployment of price-based demand response programs, power system stability becomes a serious concern [23].

TABLE I

INTERPRETATION OF PARAMETERS ϕ.

parameter	meaning
n_-, n_+	charge/discharge power limits
C	energy capacity
α	dissipation rate

II. GENERALIZED BATTERY MODELS

Generalized battery models offer a compact representation for a *set of power signals*. These models will prove useful to represent the aggregate flexibility of a collection of loads.

Definition 1: A *Generalized Battery Model* $\mathbb{B}$ is a set of signals $u(t)$ that satisfy

$$-n_- \leq u(t) \leq n_+, \quad \forall\ t > 0,$$
$$\dot{x} = -\alpha x - u,\ x(0) = 0\ \Rightarrow |x(t)| \leq C, \quad \forall\ t > 0.$$

The model is specified by the non-negative parameters $\phi = (C, n_-, n_+, \alpha)$, and we write this compactly as $\mathbb{B}(\phi)$. □

Remark 2: If we regard $u(t)$ as the power drawn from the battery and $x(t)$ as its state-of-charge at time t, the parameters ϕ have natural interpretations as summarized in Table I. For this reason, we view $\mathbb{B}$ as a generalized electricity storage. We refer to the charge/discharge rate limits n_-, n_+ as the *power limits* and C as the *energy capacity* of the battery. It will happen that the parameters ϕ are random processes that depend on ambient temperature and participation rates. As a result, we regard $\mathbb{B}(\phi)$ as a stochastic battery. □

While the state-of-charge is defined to be bounded by C, it may be more severely constrained by the bounds on power signal $u(t)$. This observation motivates the following:

Definition 3: The *effective up and down energy capacities* C_+ and C_- of the stochastic battery model $\mathbb{B}(\phi)$ are defined as

$$C_+ = \min\{C,\ n_+/\alpha\}, \quad C_- = \min\{C,\ n_-/\alpha\}. \qquad \square$$

III. INDIVIDUAL TCL MODELS

A. Dead-Band Model

We first present a standard hybrid-system model for a TCL. We will use this model for our simulation studies. In this model, the temperature evolution of a TCL is described as

$$\dot{\theta}(t) = \begin{cases} -a(\theta(t) - \theta_a) - bP_m + w, & \text{if } q(t) = 1, \\ -a(\theta(t) - \theta_a) + w, & \text{if } q(t) = 0. \end{cases} \tag{1}$$

The state of the unit is captured by the binary signal $q(t)$. We say the unit is in the ON state at time t if $q(t) = 1$, and in the OFF state if $q(t) = 0$. The unit cycles between ON and OFF states when the temperature crosses user-specified temperature thresholds:

$$\lim_{\epsilon \to 0} q(t + \epsilon) = \begin{cases} q(t), & |\theta(t) - \theta_r| < \Delta, \\ 1 - q(t), & |\theta(t) - \theta_r| \geq \Delta. \end{cases}$$

TABLE II

TYPICAL PARAMETER VALUES FOR A RESIDENTIAL AC UNIT.

Parameter	Description	Value	Unit
C_{th}	thermal capacitance	2	kWh/°C
R_{th}	thermal resistance	2	°C/kW
P_m	rated electrical power	5.6	kW
η	coefficient of performance	2.5	
θ_r	temperature setpoint	22.5	°C
Δ	temperature deadband	0.3	°C

Here, Δ is the dead-band, θ_r is the set-point, and the process noise w accounts for disturbances. The constants

$$a = \frac{1}{R_{th}C_{th}}, \quad b = \frac{\eta}{C_{th}},$$

can be expressed in terms of the thermal resistance R_{th}, thermal capacitance C_{th}, and coefficient of performance η. See [10], [11], [15] for more details on the model. The rated power P_m is positive for cooling devices, and it is negative for heating devices. Table II describes the parameters and their typical values for a residential air conditioner [24]. This simple first-order hybrid model is an approximation. The cooling dynamics of west facing houses with exposed attic spaces require higher order models. A detailed and explicit treatment of modeling uncertainty in (1) is outside the scope of this paper.

The average power consumed by a TCL over a cycle is

$$P_a = \frac{P_m T_{\text{ON}}}{T_{\text{ON}} + T_{\text{OFF}}}, \tag{2}$$

where T_{ON} is the ON state duration, and T_{OFF} is the OFF state duration per cycle. It is straightforward to show that

$$T_{\text{ON}} = R_{th}C_{th}\ \ln \frac{\theta_r + \Delta - \theta_a + R_{th}P_m\eta}{\theta_r - \Delta - \theta_a + R_{th}P_m\eta},$$
$$T_{\text{OFF}} = R_{th}C_{th}\ \ln \frac{\theta_r - \Delta - \theta_a}{\theta_r + \Delta - \theta_a}.$$

B. Continuous Power Model

As a more convenient abstraction of the dead-band model, we consider a continuous thermal model. We will use this approximate model for all our analytical work. Here, the TCL accepts any continuous power input $p(t) \in [0, P_m]$ and the dynamics are:

$$\dot{\theta}(t) = -a(\theta(t) - \theta_a) - bp(t). \tag{3}$$

Note that in this model, as common in the literature, the disturbance w is assumed to be zero mean [10], [11]. If w has non-zero mean, it could be treated by an appropriate change to θ_a. Maintaining the temperature $\theta(t)$ within the user-specified dead-band $\theta_r \pm \Delta$ is treated implicitly as a *constraint* on the power signal $p(t)$. When evaluating the trajectory $\theta(t)$ of this continuous model, it is assumed that $\theta(0) = \theta_r$. The parameters that specify this continuous power model are $\chi = (a, b, \theta_a, \theta_r, \Delta, P_m)$.

The nominal power required to keep a TCL at its set-point is given by

$$P_o = \frac{a(\theta_a - \theta_r)}{b} = \frac{\theta_a - \theta_r}{\eta R_{th}}. \tag{4}$$

Simple calculations with typical parameters reveal that the nominal power P_o under the continuous power model closely follows the average power P_a under the dead-band model (as in (2) for a wide range of operating conditions.

Remark 4: It can be shown that the continuous model is asymptotically exact for a collection of *homogeneous* TCLs represented by the hybrid model. More precisely, a load trajectory that is feasible for one is feasible for the other within a constant tolerance that decreases with the size of the collection plus a time varying tolerance that decreases to zero at least as fast as the time constant a of the TCLs. The fact that the feasible load trajectory for the collection with the hybrid model is feasible (in sum) for the continuous model follows simply by linearity. In this case, the variable θ for the continuous model is exactly the average temperature of the homogeneous TCL collection. The reverse is true because a feasible trajectory for the continuous model can be divided among a collection of hybrid units in such a way as to drive the temperatures of the hybrid units to asymptotically approach the temperature of the continuous model, and in the worst case the convergence is bounded by a term proportional to e^{-at}. For a collection of *heterogeneous* TCLs, we can cluster them into groups so that TCLs in each group have similar parameters. Each of these groups with the hybrid model can be approximated by that with the continuous model. □

IV. Aggregate Flexibility

Consider a diverse collection of TCLs indexed by k. Let P_o^k denote the nominal power consumed by the k^{th} TCL. Each TCL can accept perturbations $e^k(t)$ around its nominal power consumption that will meet user-specified comfort bounds. Define

$$\mathbb{E}^k = \left\{ e^k(t) \,\middle|\, \begin{array}{c} 0 \le P_o^k + e^k(t) \le P_m^k, \\ P_o^k + e^k(t) \text{ keeps } |\theta^k(t) - \theta_r^k| \le \Delta^k \end{array} \right\}. \quad (5)$$

This set of power signals represents the flexibility of the k^{th} TCL with respect to nominal. The *aggregate flexibility* of the collection of TCLs is defined as the Minkowski sum

$$\mathbb{U} = \sum_k \mathbb{E}^k. \quad (6)$$

The geometry of this set is, in general, unwieldy. Our objective is to develop succinct characterizations of the aggregate flexibility set. In our central result, we show that $\mathbb{U}$ can be bounded by generalized battery models as

$$\mathbb{B}(\phi_1) \subseteq \mathbb{U} \subseteq \mathbb{B}(\phi_2).$$

We have the following:

Theorem 5: Consider a collection of *heterogeneous* TCLs modeled by the continuous-power model with parameters χ^k. The aggregate flexibility $\mathbb{U}$ of the collection satisfies

$$\mathbb{U} \subseteq \mathbb{B}(\phi_2),$$

where the parameters $\phi_2 = (C, n_-, n_+, \alpha)$ are given by

$$C = \sum_k \left(1 + \left|1 - \frac{a^k}{\alpha}\right|\right) \frac{\Delta^k}{b^k},$$
$$n_- = \sum_k P_o^k, \quad n_+ = \sum_k P_m^k - P_o^k,$$

where $P_o^k = a^k(\theta_a^k - \theta_r^k)/b^k$ and $\alpha > 0$ is arbitrary. □

Proof: See Appendix A. ■

We next examine sufficient characterizations of $\mathbb{U}$. There are many choices of battery parameters ϕ such that $\mathbb{B}(\phi) \subseteq \mathbb{U}$. We have the following:

Theorem 6: Consider a collection of *heterogeneous* TCLs modeled by the continuous-power model with parameters χ^k. Fix $\alpha > 0$, and define

$$f^k = \frac{\Delta^k}{b^k \left(1 + \left|\frac{\alpha - a^k}{a^k}\right|\right)}.$$

Fix $\beta^k \ge 0, k = 1, \ldots, N$, with $\sum_k \beta^k = 1$. Let (C, n_-, n_+) be any triple that satisfies the constraints

$$\left.\begin{array}{rcl} \beta^k n_- & \le & P_o^k \\ \beta^k n_+ & \le & P_m^k - P_o^k \\ \beta^k C & \le & f^k \end{array}\right\}. \quad (7)$$

Then, the aggregate flexibility $\mathbb{U}$ of the collection satisfies

$$\mathbb{B}(\phi_1) \subseteq \mathbb{U},$$

where $\phi_1 = (C, n_-, n_+, \alpha)$.
Further, if $u \in \mathbb{B}(\phi_1)$, the causal power allocation strategy

$$e^k(t) = \beta^k u(t),$$

satisfies the dead-band constraints $|\theta^k(t) - \theta_r^k| \le \Delta^k$. □

Proof: See Appendix B. ■

Theorem 6 informs us that there are many battery models $\mathbb{B}(\phi_1)$ that offer sufficient characterizations of the aggregate flexibility set $\mathbb{U}$. In some situations we may seek a sufficient battery model with largest energy capacity C, and in others with largest charge power limit n_-, or with largest discharge power limit n_+. Table III summarizes the three extreme cases: maximize capacity C, charge rate n_-, or discharge rate n_+. The results follow from Theorem 6 by setting β^k as $\frac{f^k}{\sum_k f^k}$, $\frac{P_o^k}{\sum_k P_o^k}$ and $\frac{P_m^k - P_o^k}{\sum_k (P_m^k - P_o^k)}$, respectively.
Since $\mathbb{B}(\phi_1) \subseteq \mathbb{U} \subseteq \mathbb{B}(\phi_2)$, there is, in general, a gap between our necessary and sufficient characterizations of aggregate flexibility. We explore the conservatism in our battery models using simulations in Section VI. For a population of *homogeneous* TCLs, our battery model characterizations are exact. More precisely, we have:

Corollary 7: Consider a collection of N *homogeneous* TCLs, modeled by the continuous-power model with parameters $\chi = (a, b, \theta_a, \theta_r, \Delta, P_m)$. Then,

$$\mathbb{U} = \mathbb{B}(C, n_-, n_+, \alpha),$$

TABLE III

COMPARISON OF GENERALIZED BATTERY MODELS FOR A COLLECTION OF HETEROGENEOUS TCLS

	(Necessary) Battery $\mathbb{B}(\phi_2)$	(Sufficient) Battery $\mathbb{B}(\phi_1)$		
		Maximize C	Maximize n_-	Maximize n_+
C	$\sum_k (1+\lvert 1-\frac{a^k}{\alpha}\rvert)\frac{\Delta^k}{b^k}$	$\sum_k f^k$	$(\sum_k P_o^k)\min_k \frac{f^k}{P_o^k}$	$(\sum_k P_m^k - P_o^k)\min_k \frac{f^k}{P_m^k - P_o^k}$
n_-	$\sum_k P_o^k$	$(\sum_k f^k)\min_k \frac{P_o^k}{f^k}$	$\sum_k P_o^k$	$(\sum_k P_m^k - P_o^k)\min_k \frac{P_o^k}{P_m^k - P_o^k}$
n_+	$\sum_k P_m^k - P_o^k$	$(\sum_k f^k)\min_k \frac{P_m^k - P_o^k}{f^k}$	$(\sum_k P_o^k)\min_k \frac{P_m^k - P_o^k}{P_o^k}$	$\sum_k P_m^k - P_o^k$

where

$$C = N\Delta/b,\ n_- = NP_o,\ n_+ = N(P_m - P_o),\ \alpha = a,$$

and $P_o = a(\theta_a - \theta_r)/b$. □

Proof: Follows immediately from Theorems 5 and 6 by setting $\alpha = a^k$. ■

Remark 8: The gap between the battery models $\mathbb{B}(\phi_2)$ and $\mathbb{B}(\phi_1)$ increases with TCL heterogeneity. To obtain tighter models, we can *cluster* TCLs with similar parameters into g groups, and compute a battery model for each group. The aggregate flexibility is then represented as the union of g generalized battery models. Theorem 6 guides the choice of metric for clustering TCLs. □

V. CONTROL ARCHITECTURE

Our objective is to reliably deliver frequency regulation service to the grid by actively controlling an aggregation of TCLs. The regulation signal or AGC (Automatic Generation Control) command $r(t)$ is typically a command determined by the system operator at 4-second sampling based on the area control error [1].

We adopt a centralized control architecture. This choice is dictated by the reliability requirements necessary to participate in regulation ancillary service markets. At each sample time, the aggregator compares the regulation signal $r(t)$ with the aggregate power deviation $\delta(t) = P_{\text{agg}}(t) - n(t)$. Here $P_{\text{agg}}(t) = \sum_k q^k(t)P_m^k$ is the instantaneous power drawn by the TCLs, and $n(t) = \sum_k P_o^k$ is their baseline power. This requires a contractual *ex ante* agreement on what the "baseline" power consumption $n(t)$ of the aggregation will be on the forward delivery window.

If $r(t) < \delta(t)$, the population of TCLs needs to "discharge" power to the grid, which means some of the ON units will be turned OFF. Conversely, if $r(t) > \delta(t)$, then the population of TCLs must consume more power. This requires turning ON some of the OFF units. The selection of the most appropriate TCLs that must be turned ON or OFF is done through a priority-stack-based control strategy which will be described in the next section. We stress that this is a *feedback* control strategy which offers robustness against disturbances w due to occupancy patterns, solar radiation, etc. and modeling errors in the dynamics of the TCLs. The overall control architecture is depicted in Fig. 1.

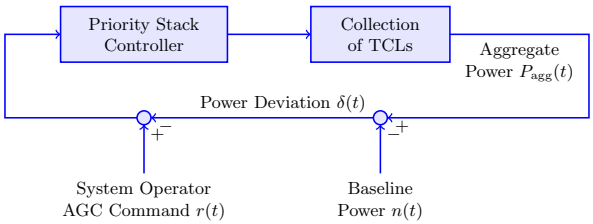


Fig. 1. Control architecture for regulation service provisoin.

A. Priority Stacks

Suppose at time t we have $r(t) < \delta(t)$. The population of TCLs must reduce their power consumption. We must therefore turn OFF some units that are ON. It is most appropriate to turn OFF those units that will most imminently turn OFF. Imminence can be measured naturally by *temperature distance* to the switching boundary, i.e. by $\pi^k(t) = (\theta^k(t) - \underline{\theta}^k)/\Delta^k$, where $\underline{\theta}^k = \theta_r^k - \Delta^k$. The temperature distance is normalized to account for heterogeneity. We can therefore construct the ON *priority stack* which consists of units that are ON. The TCLs in this stack are ordered by their priority criterion, $\pi^k(t)$. Analogously, we construct the OFF priority stack of units that are OFF, ordered by $\pi^k(t) = (\overline{\theta}^k - \theta^k(t))/\Delta^k$, where $\overline{\theta}^k = \theta_r^k + \Delta^k$.

Imminence can also be measured by *time* to the switching boundary. For example, in the ON priority stack, units that will turn OFF autonomously the soonest receive the highest priority.

The unit with the highest priority will be turned ON (or OFF) first, and then units with lower priorities will be considered in sequence until the desired regulation is achieved. This

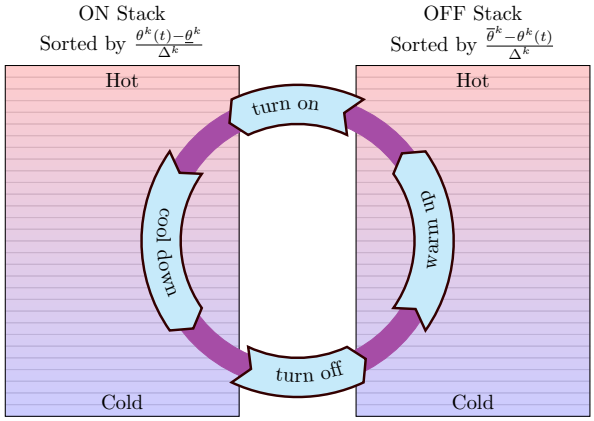


Fig. 2. The ON and OFF priority stacks.

priority-stack-based control strategy minimizes the ON/OFF switching action for each unit, which avoids short-cycling and reduces wear and tear of the mechanical equipment. Priority stacks are illustrated in Fig. 2. We index the units available for manipulation in the ON stack from *bottom to top* by $\{1, 2, \cdots, N_1\}$, and the units available for manipulation in the OFF stack from *top to bottom* by $\{1, 2, \cdots, N_0\}$. Units are turned ON or OFF when their *real time* power consumption $P^k(t)$ matches the difference between the regulation signal $r(t)$ and aggregate power deviation $\delta(t)$. The associated control algorithm is summarized in Algorithm 1.

Algorithm 1 Priority-stack-based control algorithm

loop
  receive $\pi^k(t)$ and $P^k(t)$;
  construct priority stacks;
  read $r(t)$;
  compute $\delta(t) = P_{\text{agg}}(t) - n(t)$;
  if $\delta(t) < r(t)$ **then**
    find $j^* = \min\left\{j \mid \sum_{i=1}^{j} P^i(t) \geq r(t) - \delta(t)\right\}$;
    turn ON units indexed by $\{1, 2, \cdots, j^*\}$.
  else if $\delta(t) > r(t)$ **then**
    find $j^* = \min\left\{j \mid \sum_{i=1}^{j} P^i(t) \geq \delta(t) - r(t)\right\}$;
    turn OFF units indexed by $\{1, 2, \cdots, j^*\}$.
  end if
end loop

Remark 9: The priority-stack-based control offers a generic architecture suitable for direct load control of various classes of flexible loads. For example, common scheduling strategies for electric vehicle fleet such as Earliest-Deadline-First define a priority stack [25]. Priorities can also be used to encode pricing and service quality: consumers who wish to exercise greater control over their loads could opt to receive low compensation when called to supply regulation services in exchange for a lower priority position. □

B. Practical Considerations

To implement the proposed direct load control strategy, we require (at a minimum) measurements of power $P^k(t)$ and temperature $\theta^k(t)$ at a sampling rate of 0.25 Hz for each TCL. While $\theta^k(t)$ and the set-point θ^k_r are directly available from the thermostat, measuring the power $P^k(t)$ requires additional hardware infrastructure. For each TCL, run-time system identification algorithms can be used to estimate the operating state $q^k(t)$, the ambient temperature θ^k_a, and model parameters a^k, b^k, Δ^k from the temperature time series $\theta^k(s), s \leq t$. Using this information, a local embedded controller computes $\pi^k(t)$ for each TCL.

The priority criterion $\pi^k(t)$ and power consumption $P^k(t)$ are transmitted to the aggregator. The aggregator forms the priority stack from the collated data and computes the control action for the next sample. This is broadcast to the TCLs where the local controller implements the action. Latency in the control loop will determine the quality of the offered regulation service in terms of power ramp rates. This scheme has modest computation and communication overhead.

Remark 10: Measuring the power consumption of each TCL necessitates nontrivial capital cost as power meters are expensive. *In our view, this is unavoidable.* Alternate and realistic schemes, Scenarios 1 and 2 of [15], have been proposed that use population-bin models and requires measuring only the TCL ON/OFF state. This requires moderately simpler sensing infrastructure, but requires a more complex control strategy. Other schemes have been proposed where the aggregate power $P_{\text{agg}}(t)$ is *estimated* using population models [16], or disaggregated from substation measurements [10], and Scenarios 3 and 4 of [15]. The former scheme is open-loop in character. The latter schemes must contend with the fact that participating TCLs represent a small fraction of the connected loads at a substation, making it very difficult to infer their power consumption from an aggregate measurement. These schemes therefore face big challenges in meeting the reliability requirements necessary to participate in the regulation ancillary service market [26]. □

Remark 11: Our scheme (as well as Scenarios 1 and 2 of [15]) requires real-time telemetry to transmit power measurements to the aggregator. This requires very little bandwidth and is becoming increasingly inexpensive. Since telemetry costs are dominated by sensing infrastructure costs, transmission represents a minor expense. Real-time telemetry is not always needed when a conventional generator resource is providing regulation. The key difference is that with TCLs we have *distributed* resources. Telemetry is needed to close the loop, i.e. send measurements to the controller which is implemented by the remote aggregator. For generators, the control is local: sensing and actuation are at the same place, so real-time telemetry may not be needed. □

VI. Simulation Studies

We consider a population of 1000 diverse air conditioners (ACs). In our simulations, we use the more accurate hybrid model (1) for each TCL. The nominal model parameters are listed in Table II. For a heterogeneous collection, we assume that the TCL parameters are drawn from a uniform distribution with 10% heterogeneity around their nominal values. For example, $R^k_{th} \sim U(0.95R_o, 1.05R_o)$ where R_o is the nominal value of the thermal resistance. The ambient temperature is assumed to be $32°C$, and the initial temperatures and operating states of the population of TCLs are randomized.

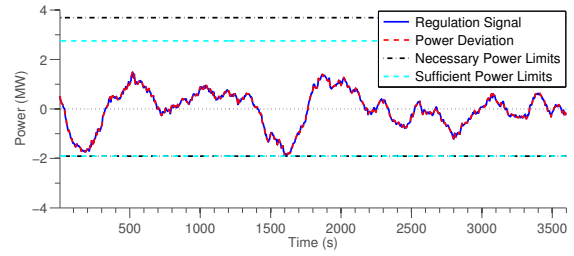
The power limits and energy capacities of the collection of TCLs using the (necessary) battery model $\mathbb{B}(\phi_2)$ and the (sufficient) battery model $\mathbb{B}(\phi_1)$ are listed in Table IV. In both models, the dissipation parameter α is given as the average of the time constants of individual TCLs. Formally, we assume $\alpha := \frac{1}{N}\sum_{k=1}^{N} 1/(R^k_{th}C^k_{th})$. For the sufficient battery model, we maximize the charge rate n_- (the fourth column in Table III). Note that the battery parameters are derived based on the continuous power model (3).

We apply our priority-stack-based control strategy to

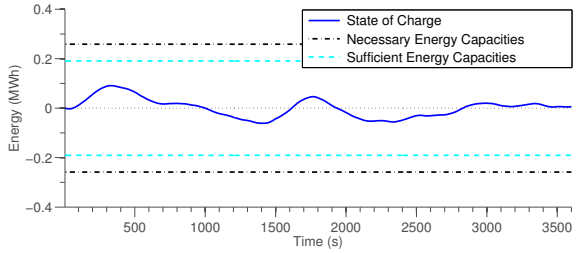
TABLE IV

POWER LIMITS AND ENERGY CAPACITIES.

	Sufficient Battery $\mathbb{B}(\phi_1)$	Necessary Battery $\mathbb{B}(\phi_2)$
n_-	1.9 MW	1.9 MW
n_+	2.8 MW	3.7 MW
C	0.19 MWh	0.26 MWh



(a) Power trajectory



(b) State of charge

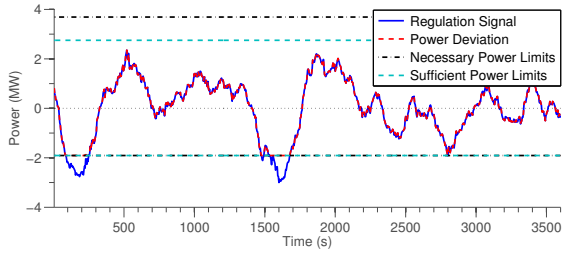
Fig. 3. Tracking of a regulation signal succeeds when it is within the power limits and energy capacity of the (sufficient) battery model $\mathbb{B}(\phi_1)$.

track a one-hour long regulation signal $r(t)$ from PJM (Pennsylvania-New Jersey-Maryland Interconnection) [27]. The magnitude of the PJM signal is scaled appropriately to match the power limits and energy capacity of 1000 ACs.

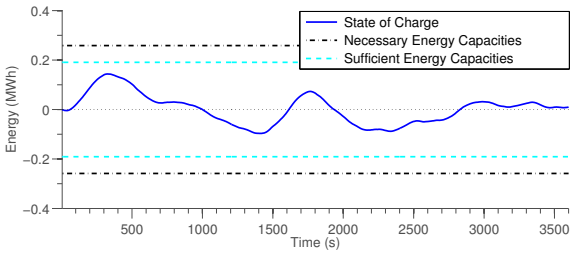
A. Tracking Performance

Fig. 3 shows that if the regulation signal $r(t)$ has both power and capacity requirements within the analytic bounds of the (sufficient) battery model $\mathbb{B}(\phi_1)$, the population of TCLs delivers excellent tracking. The maximum tracking error is less than 1% of the maximum magnitude of the regulation signal. Additional simulation results (not reported here) reveal that even with one sample (4 sec) communication delay, good tracking is still achieved with a maximum tracking error less than 5% of the maximum magnitude of the regulation signal. If the regulation signal violates either the power limits or energy capacity of the (necessary) battery model $\mathbb{B}(\phi_2)$, the population of TCLs fails to track the regulation signal. Figs. 4 and 5 show that when the regulation signal exceeds the power limits or the energy capacity respectively, we cannot track the regulation signal. Extensive simulations (not reported here) using other regulation signals yield similar conclusions.

We use a typical 6-hour long regulation signal from PJM (shown in Fig. 6 (a)) that is fairly close to the power limits and energy capacity to test the prediction performance of our sufficient battery model. Specifically, we examine the effect of tracking of a regulation signal that (just) satisfies the sufficient battery model on the number of additional ON/OFF

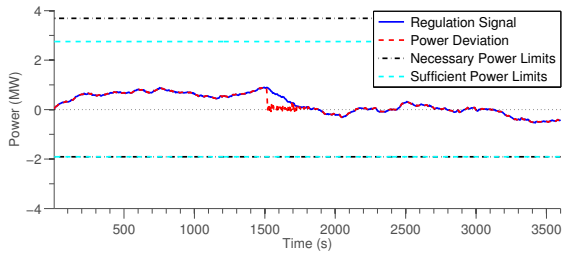


(a) Power trajectory

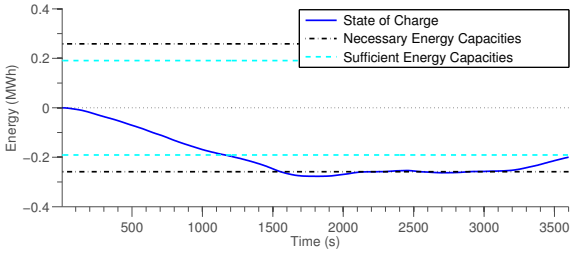


(b) State of charge

Fig. 4. Tracking of a regulation signal fails when it exceeds the power limit of the (necessary) battery model $\mathbb{B}(\phi_2)$.



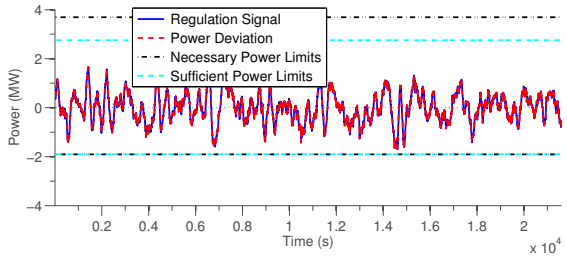
(a) Power trajectory



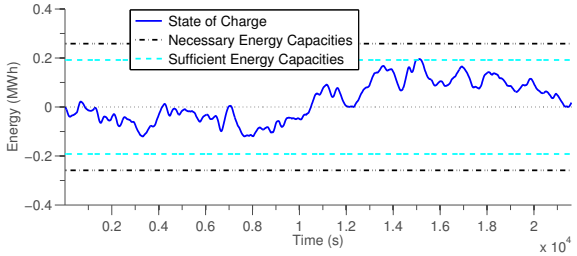
(b) State of charge

Fig. 5. Tracking of a regulation signal fails when it exceeds the energy capacity of the (necessary) battery model $\mathbb{B}(\phi_2)$.

switchings that occur above nominal, and occurrence of short cycling events. Table V shows the performance statistics. The average number of ON/OFF switching in the 6-hour period without providing regulation service (nominal) is 30 times, while the average switching in the same period with regulation using the priority-stack controller is 65 times, or about twice of the nominal value. Additionally, we assume for each unit, its minimum ON-time and OFF-time are 1 minute. With provision of frequency regulation, the minimum, average, and maximum number of short cycling events among all units in the 6-hour period are respectively 0, 1.1, and 8. Extensive simulations using other regulation signals are also conducted, and similar statistics are obtained. We observe that the number of short cycling events is relatively small,



(a) Power trajectory



(b) State of charge

Fig. 6. Tracking of a typical 6-hour regulation signal from PJM that is within the power limits and energy capacity of the (sufficient) battery model $\mathbb{B}(\phi_1)$.

TABLE V
PREDICTION PERFORMANCE OF THE SUFFICIENT BATTERY MODEL

	Switchings		Short Cyclings		
	nominal	regulation	min	mean	max
Sufficient Battery	30	65	0	1.1	8

and would be drastically reduced with a less aggressive regulation signal. Note also that short cycling events are closely connected to large, high frequency oscillations in the regulation signal, and an exact characterization of feasibility with a short cycling constraint is reported in [28].

B. Battery Model Conservatism

Recall that the gap between the necessary and sufficient battery models is due to heterogeneity in the collection of TCLs (see Corollary 7). We synthetically vary the diversity d in the TCL parameters from 1% to 40%. In each case, we then scale the AGC signal $r(t)$ until our control system fails to provide tracking on a 1-hour window. More precisely, we numerically compute (for each d),

$$\max \gamma: \quad \text{tracking error } \leq 1\% \text{ for AGC command } \gamma r(t)$$

We declare the "true" battery power limits and energy capacity to be

$$n_-^{\text{true}} = -\gamma \min_t r(t), \quad n_+^{\text{true}} = \gamma \max_t r(t),$$
$$C^{\text{true}} = \max |x(t)|, \quad \dot{x} = -\alpha x - \gamma r(t), \; x(0) = 0.$$

Fig. 7 compares the numerical energy capacity to the bounds given by the necessary and sufficient battery models as a function of diversity d. We note that for $d < 10\%$, the models capture the aggregate flexibility quite well. Similar plots can be obtained for power limits also.

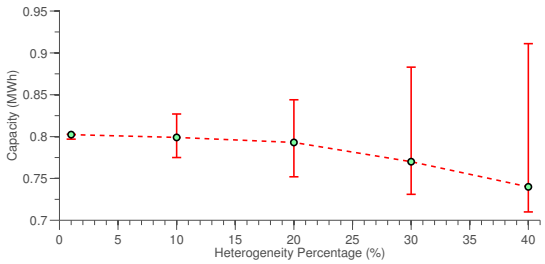


Fig. 7. Conservatism in the bounds on energy capacity from the necessary and sufficient models. The green dots show the "true" energy capacity as calculated numerically. The red bars extend between the energy capacity bounds given by the sufficient and necessary battery models.

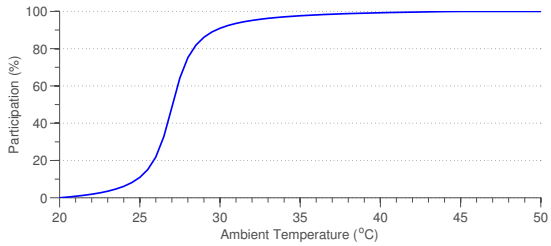


Fig. 8. Participation percentage vs. ambient temperature.

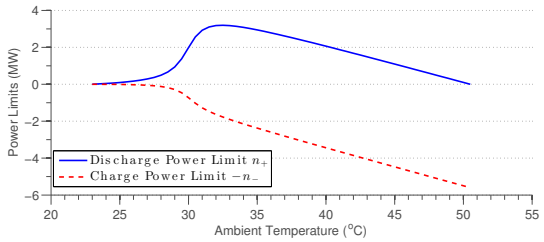
C. The Effect of Ambient Temperature

We have assumed that all of the available TCLs participate in offering regulation services. In practice, for certain types of TCLs such as air conditioners and heat pumps, participation rates depend strongly on ambient temperature θ_a. For air conditioners, we expect little participation when θ_a is low, and significant participation when θ_a is high. Modeling participation is extremely complex, requires large amounts of data, and any resulting models will likely have limited predictive power. For purposes of illustration, suppose we synthetically model participation using an inverse tangent function as shown in Fig. 8. This captures the intuitive observation that more people are likely to use their AC units at higher ambient temperatures. This does not account for occupancy which exhibits daily and hourly patterns. With this participation model, we can compute the sufficient battery parameters ϕ_1 as a function of ambient temperature. These are shown in Fig. 9 for air-conditioning loads.

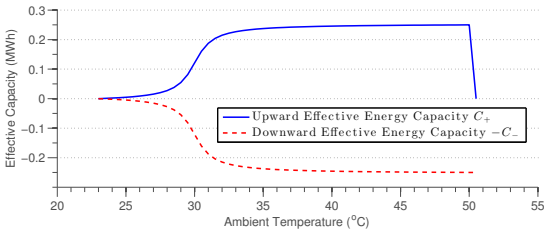
VII. Conclusions and Future Work

In this paper, we illustrated that (a) the generalized battery model provided a succinct and powerful framework to characterize the aggregate flexibility of a population of TCLs, (b) the power and capacity bounds derived from the continuous model accurately captured the aggregate flexibility of TCLs with the hybrid model, and (c) the priority-stack-based control strategy yielded excellent tracking performance and good robustness.

The enormous potential of TCLs presents a tremendous opportunity for providing regulation service to the grid. There are several important research issues that must be addressed to realize this vision. These include: (a) deriving



(a) Power limits



(b) Effective energy capacities

Fig. 9. Power limits and effective capacities vs. ambient temperature.

battery models that account for TCL model uncertainty, (b) an exploration of suitable and low-cost firmware and communication infrastructure to implement direct-load control, (c) understanding the quality of regulation service provided in terms of latency and ramp rates, (d) estimating the overall hourly availability of TCLs using historic measurement data, and (e) developing fair schemes to compensate loads for participating in regulation services.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Pravin Varaiya and Laurent Lessard for insightful comments and discussions.

APPENDIX

For a signal x with support $[0,\infty)$, we define the standard norms $\|x\|_\infty = \max_t |x(t)|, t \geq 0$ and $\|x\|_1 = \int_0^\infty |x(t)|dt$.

A. Proof of Theorem 5

Let u be any element of $\mathbb{U}$. Then $u = \sum_k e^k$, where $e^k \in \mathbb{E}^k$. The proof requires showing that $u \in \mathbb{B}(\phi_2)$. Define $\vartheta^k(t) = (\theta^k(t) - \theta_r^k)/b^k$. Since $\theta^k(0) = \theta_r^k$, we have from (3),

$$\dot{\vartheta}^k(t) = -a^k\vartheta^k(t) - e^k(t), \quad \vartheta^k(0) = 0.$$

Taking a Laplace transform of the above equation, we obtain

$$\vartheta^k(s) = -\frac{1}{s+a^k}e^k(s).$$

Let $\dot{x}(t) = -\alpha x(t) - u(t)$ and $x(0) = 0$. Taking the Laplace transform,

$$\begin{aligned} x(s) &= -\frac{1}{s+\alpha}u(s) = -\frac{1}{s+\alpha}\sum_k e^k(s) \\ &= -\sum_k \frac{1}{s+\alpha}e^k(s) = -\sum_k \frac{s+a^k}{s+\alpha}\frac{1}{s+a^k}e^k(s) \\ &= \sum_k \frac{s+a^k}{s+\alpha}\vartheta^k(s) = \sum_k (1+\frac{a^k-\alpha}{s+\alpha})\vartheta^k(s). \end{aligned}$$

Given $Y(s) = H(s)U(s)$, we have $\|y(t)\|_\infty \leq \|h(t)\|_1\|u(t)\|_\infty$, where $y(t), h(t)$ and $u(t)$ are the inverse Laplace transforms of $Y(s), H(s)$ and $U(s)$ respectively. Therefore,

$$\|x(t)\|_\infty \leq \sum_k (1+|1-\frac{a^k}{\alpha}|)\|\vartheta^k(t)\|_\infty.$$

Because $e^k(t) \in \mathbb{E}^k$, $\|\vartheta^k(t)\|_\infty \leq \Delta^k/b^k$, so that

$$\|x(t)\|_\infty \leq \sum_k (1+|1-\frac{a^k}{\alpha}|)\Delta^k/b^k. \tag{8}$$

Additionally, $-P_o^k \leq e^k(t) \leq P_m^k - P_o^k$, implies that

$$-\sum_k P_o^k \leq u(t) \leq \sum_k P_m^k - P_o^k. \tag{9}$$

Inequalities (8) and (9) verify that $u(t) \in \mathbb{B}(\phi_2)$. ■

B. Proof of Theorem 6

In this case, it is sufficient to show that $u(t) \in \mathbb{B}(\phi_1)$ implies $e^k(t) = \beta^k u(t) \in \mathbb{E}^k$. Let $\dot{x} = -\alpha x - u, x(0) = 0$. Now, if $u(t) \in \mathbb{B}(\phi_1)$, the following must hold

$$\begin{aligned} -n_- \leq u(t) \leq n_+, \qquad & \forall\ t > 0, \\ |x(t)| \leq C, \qquad & \forall\ t > 0. \end{aligned}$$

Apply power deviation $e^k(t) = \beta^k u(t)$ to each TCL unit. Define $\vartheta^k(t) = (\theta^k(t) - \theta_r^k)/b^k$ with $\theta^k(0) = \theta_r^k$. From (3),

$$\dot{\vartheta}^k(t) = -a^k\vartheta^k(t) - e^k(t), \quad \vartheta^k(0) = 0.$$

Taking the Laplace transform of the above equation,

$$\begin{aligned} \vartheta^k(s) &= -\frac{1}{s+a^k}e^k(s) = -\frac{1}{s+a^k}\beta^k u(s) \\ &= -\beta^k\frac{s+\alpha}{s+a^k}\frac{u(s)}{s+\alpha} = \beta^k(1+\frac{\alpha-a^k}{s+a^k})x(s), \end{aligned}$$

where we have used the fact that $x(s) = -\frac{1}{s+\alpha}u(s)$. Using the same bounds as in the proof of Theorem 5,

$$\begin{aligned} \|\vartheta^k(t)\|_\infty &\leq \beta^k(1+|\frac{\alpha-a^k}{a^k}|)\|x(t)\|_\infty \\ &\leq \beta^k(1+|\frac{\alpha-a^k}{a^k}|)C, \quad \forall\ t > 0. \end{aligned}$$

Since C is chosen so that $\beta^k C \leq \frac{\Delta^k}{b^k\left(1+\left|\frac{\alpha-a^k}{a^k}\right|\right)}$, we have $|\vartheta^k(t)| \leq \Delta^k/b^k$ or equivalently $|\theta^k(t) - \theta_r^k| \leq \Delta^k$. Moreover $\beta^k n_- \leq P_o^k$ and $\beta^k n_+ \leq P_m^k - P_o^k, \forall\ k$, yields $-P_o^k \leq \beta^k u(t) \leq P_m^k - P_o^k$ which shows that $e^k(t) = \beta^k u(t) \in \mathbb{E}^k$. ■

REFERENCES

[1] B. Kirby, "Ancillary services: Technical and commercial insights," Tech. Rep., July 2007.

[2] J. Smith, M. Milligan, E. 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, aug. 2007.

[3] Y. 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, may 2009.

[4] S. Meyn, M. Negrete-Pincetic, G. Wang, A. Kowli, and E. Shafieepoorfard, "The value of volatile resources in electricity markets," in *IEEE conference on Decision and Control (CDC)*, 2010.

[5] 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*, 2010.

[6] Pacific Gas and Electric Company, "Smart AC program." [Online]. Available: http://www.pge.com/en/myhome/saveenergymoney/energysavingprograms/smartac/index.page

[7] "Buildings energy data book." [Online]. Available: http://buildingsdatabook.eren.doe.gov/default.aspx

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

[9] H. Hao, B. Sanandaji, K. Poolla, and T. Vincent, "A generalized battery model of a collection of thermostatically controlled loads for providing ancillary service," in *the 51th Annual Allerton Conference on Communication, Control and Computing*, 2013.

[10] 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.

[11] 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.

[12] S. Kundu, N. Sinitsyn, S. Backhaus, and I. Hiskens, "Modeling and control of thermostatically controlled loads," in *the 17-th Power Systems Computation Conference*, 2011.

[13] S. Bashash and H. K. Fathy, "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.

[14] S. Shao, M. Pipattanasomporn, and S. Rahman, "Development of physical-based demand response-enabled residential load models," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 543–550, 2012.

[15] 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, Feburary 2013.

[16] 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*, 2013.

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

[18] 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.

[19] E. Vrettos, S. Koch, and G. Andersson, "Load frequency control by aggregations of thermally stratified electric water heaters," in *3rd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies*. IEEE, October 2012, pp. 1–8.

[20] H. Chao, "Price-responsive demand management for a smart grid world," *The Electricity Journal*, vol. 23, no. 1, pp. 7–20, 2010.

[21] H. Shu, R. Yu, and S. Rahardja, "Dynamic incentive strategy for voluntary demand response based on TDP scheme," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–6.

[22] NERC, "Data collection for demand side management for quantifying its influence on reliability: Results and recommendations," Tech. Rep., December 2007. [Online]. Available: http://www.nerc.com/docs/pc/drdtf/NERC_DSMTF_Report_040308.pdf

[23] M. Roozbehani, M. A. Dahleh, and S. K. Mitter, "Volatility of power grids under real-time pricing," *IEEE Transactions on Power Systems*, 2011.

[24] J. Mathieu, M. Dyson, and D. Callaway, "Using residential electric loads for fast demand response: The potential resource and revenues, the costs, and policy recommendations," in *2012 ACEEE Summer Study on Energy Efficiency in Buildings*, 2012.

[25] A. Subramanian, M. Garcia, A. Dominguez-Garcia, D. Callaway, K. Poolla, and P. Varaiya, "Real-time scheduling of deferrable electric loads," in *American Control Conference (ACC)*, 2012, pp. 3643–3650.

[26] J. MacDonald, P. Cappers, D. Callaway, and S. Kiliccote, "Demand response providing ancillary services," in *Grid-Interop 2012*, Irving, Texas, December 2012.

[27] "PJM regulation data." [Online]. Available: http://www.pjm.com/markets-and-operations/ancillary-services/mkt-based-regulation/fast-response-regulation-signal.aspx

[28] B. M. Sanandaji, H. Hao, K. Poolla, and T. L. Vincent, "Improved battery models of an aggregation of thermostatically controlled loads for frequency regulation," *to appear in American Control Conference*, June 2014. [Online]. Available: http://www.eecs.berkeley.edu/~sanandaji/my_papers.html