# Fast Regulation Service Provision via Aggregation of Thermostatically Controlled Loads

Borhan M. Sanandaji*
*UC Berkeley*
*Berkeley, U.S.A.*
*sanandaji@berkeley.edu*

He Hao
*UC Berkeley*
*Berkeley, U.S.A.*
*hehao@berkeley.edu*

Kameshwar Poolla
*UC Berkeley*
*Berkeley, U.S.A.*
*poolla@berkeley.edu*

***Abstract*—Federal Energy Regulatory Commission (FERC) Order 755 requires scheduling coordinators to procure and compensate more for regulation resources with faster ramping rates. Thermostatically Controlled Loads (TCLs) are a tremendous demand-side resource for providing fast regulation service due to their population size and their ability of being turned ON or OFF simultaneously. In this paper, we consider modeling and control of a collection of TCLs to provide such regulation service. We first develop a non-uniform bin state transition model for aggregate modeling of a collection of TCLs. The non-uniform model presents a potential for more accurate prediction while requiring fewer number of bins (reducing the complexity of the model) than the existing uniform bin models. We also propose a randomized priority control strategy to manipulate the power consumption of TCLs to track a regulation signal, while preventing short cycling, and reducing wear and tear on the equipment. The proposed control strategy is decentralized in the sense that each TCL makes its own decision solely based on a common broadcast command signal. This framework reduces the communication and computational efforts required for implementing the control strategy. We provide illustrative simulations to show the accuracy of the developed non-uniform model and efficacy of the proposed control strategy.**

***Keywords*-Fast Ancillary Service; Thermostatically Controlled Loads; Bin State Model; Randomized Priority Control**

## I. Introduction

### A. The Need for Fast Demand-Side Regulation Resources

The envisioned future grid requires much greater penetration of renewable resources than the current level. However, volatility, stochasticity, and intermittency characteristics of renewable energies present a challenge for integrating these resources into the existing grid in a large scale as the proper functioning of an electric grid requires a continuous power balance between supply and demand. To ensure the reliability of the grid, more ancillary service procurements from generations and flexible loads are required.

*Corresponding author. This work was supported in part by EPRI and CERTS under sub-award 09-206; PSERC S-52; NSF under Grants 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.

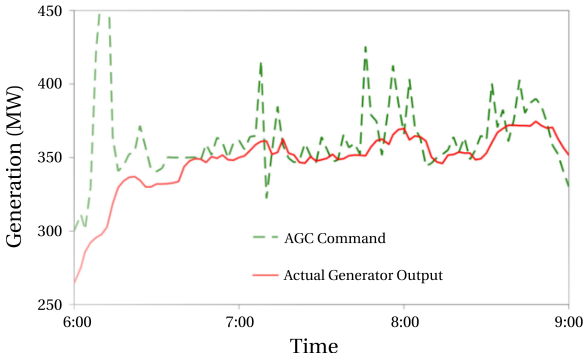



Figure 1. A coal-fired power generator follows AGC commands poorly [1].

Regulating reserve is one of the most important ancillary services for maintaining the power balance in normal conditions [1]. It is deployed in seconds (up to one minute) time scales to compensate for the short term fluctuations in the total system load and uncontrolled generation. This service has been traditionally provided by generators. However, traditional generators have slow ramping rates and cannot track the fast changing regulation signal very well. As an example, Fig. 1 illustrates how a coal-fired[1] power generator fails to follow the Automatic Generation Control (AGC) setpoint commands closely [1]. This issue has been recognized in the power and energy community [1], [2]. In particular, the Federal Energy Regulatory Commission (FERC) order 755 has been issued to recommend the system operators to pay more for faster regulation resources.

Studies show that increased reliance on renewables imposes additional regulation requirements on the grid [3]–[5]. For instance, it is shown that if California adopts its 33% renewable penetration target by 2020, the regulation procurement is anticipated to increase from 0.6 GW to 1.3 GW [6], [7]. If these ancillary service procurements are provided by fossil fuel generators, it will be counterproductive to carbon emission reductions and will be economically untenable.

The regulation requirements can be lowered if faster responding resources are available [8]. It was shown that if California Independent System Operator (CAISO) dis-

[1]Similar characteristics are observed for other thermal generators.

patched fast responding regulation resources, it could reduce its regulation procurement by as much as $40\%$ [9]. Moreover, the Midwest Independent System Operator (MISO) has reported a reduction in the total regulation payment and an improvement on the Control Performance Standard (CPS) since it has increased the share of fast ramping resources in the regulation market [10]. These factors coupled with the search for cleaner sources of flexibility as well as regulatory developments such as FERC order 755 have motivated a growing interest in tapping fast responding demand-side resources.

### B. Related Work

Buildings in the United States account for $75\%$ of the total electricity consumption with approximately equal shares between residential and commercial buildings. Buildings are hence natural candidates for providing demand-side flexibility. Related prior work on buildings has shown that fast regulation service can be provided by using variable frequency drive to control the supply fan in the HVAC system of commercial buildings [11]. It was reported that $15\%$ of rated fan power can be tapped for regulation service with little impact on the indoor environment. In this paper, we argue that residential Thermostatically Controlled Loads (TCLs) present a larger potential than commercial buildings to provide fast regulation service. This is due to the fact that with the same rated power as commercial buildings, a population of TCLs has the ability to be turned ON or OFF simultaneously, resulting in larger potential and faster ramping rate.

The thermal storage potential of TCLs was recognized as early as the 1980s [12]–[14]. In fact, many utility companies have already harnessed the flexibility of residential loads for demand response. For example, the SmartAC$^{\text{TM}}$ program initiated by Pacific Gas and Electric Company (PG&E) gathered $147,600$ residential customers for peak load shaving and managing emergency situations [15]. In addition, Florida Power and Light Company (FPL) has $780,000$ customers enrolled in their OnCall® program in which residential AC, water heaters, and pool pumps are centrally controlled in response to grid requirements [16]. However, these load control mechanisms implemented are primarily concerned with *low frequency* changes of demand (i.e., changes that occur over a timescale of hours). More recent approaches have proposed aggregate modeling and control of residential TCLs for ancillary service and energy arbitrage [17]–[22]. In particular, Malhame and Chong [14] used a statistical approach for modeling a population of TCLs as a system of coupled Fokker-Planck equations. Another approach is by considering a bin state transition model for aggregate modeling of a population of TCLs [19]–[21]. There is an enormous additional potential, particularly for fast regulation service, that is virtually not exploited.

### C. Summary of Contributions

In this paper, we focus on high frequency load changes for providing fast regulation service. We first study aggregate modeling of a population of residential TCLs. To this end, we propose a non-uniform bin state transition model, which is an extension of the uniform bin models proposed in [19]–[21]. In this setting, the length of each bin is different. The motivation behind this framework is to have the same transition rate from one bin to the next level bin for each pair of neighboring bins. In our state transition model, we further allow different number of bins for the ON and OFF states. We demonstrate that a non-uniform partitioning of the temperature deadband increases the accuracy of the prediction model and even with a smaller number of bins.

We next consider a "randomized priority control" framework for manipulating TCLs. In this framework, the controller gives a higher priority to the units that are going to be naturally switched by their local control law. In addition, this control strategy is decentralized as each unit makes its control decision based on its own temperature and a common broadcast switching probability. This control framework also prevents short cycling (which happens when a TCL is switched ON or OFF quicker than allowed by the manufacturer) and reduce wear and tear on TCLs.

The remainder of this paper is organized as follows. Section II and Section III respectively describe the individual and aggregated TCL models. In Section IV, we propose a decentralized priority control framework to control TCLs for fast regulation service provision. Numerical simulations and findings are reported in Section V. The paper ends with conclusions and future work summarized in Section VI.

### D. Nomenclature

Before getting into the details, we list and define some of the terms that will be frequently used throughout the paper. The terms that are not mentioned in this section will be explained and defined in details as needed throughout the paper. Table I summarizes some of these terms.

## II. Individual TCL Model

### A. Continuous-Time Model

Consider a population of $N$ TCLs. The temperature evolution of the $i$-th TCL can be described as

$$C^i \frac{d\theta^i(t)}{dt} = \frac{\theta_a - \theta^i(t)}{R^i} - m^i(t) P^i \eta^i + w^i(t), \qquad (1)$$

where $\theta^i(t)$ is the internal temperature of the $i$-th TCL, and $C^i$ and $R^i$ are respectively its thermal capacitance and thermal resistance, $\theta_a$ is the ambient temperature, and $m^i$ is a dimensionless binary variable that indicates the operating state of each TCL (i.e., $1$ when it is ON and $0$ when it is OFF). In addition, $P^i$ is the rated power of each unit and $\eta^i$ is its coefficient of performance. The first term on the right-hand side of (1), $(\theta_a - \theta^i(t))/R^i$, represents the heat

Table I
NOMENCLATURE OF SOME OF THE FREQUENTLY-USED TERMS.

| Term | Description |
|---|---|
| Temperature Bin | The temperature deadband can be divided to $N$ uniformly- or non-uniformly-spaced slots called temperature bins. |
| Bin State | We associate 2 states corresponding to the number of ON and OFF units in a given temperature bin. |
| Homogeneous Collection | A collection of TCLs is called homogeneous when their model parameters as given in Table II are the same. |
| Heterogeneous Collection | A collection of TCLs is called heterogeneous when their model parameters as given in Table II are different. |
| Regulating Reserve | It is one of the key ancillary services for normal conditions. It is considered to respond rapidly (seconds to one minute time scale) to compensate for fluctuations in system load and generation. |
| Load Following | Similar to regulating reserve, load following is also one of key ancillary services in normal conditions. However, it has been deployed on slower time scales (up to several minutes). |
| Regulation Signal | An AGC signal that contains system operator's setpoint commands at every 4 seconds. This signal is constructed in order to maintain the power system's frequency at the desired level by balancing control area's generation and load. |
| Short Cycling | It happens when a TCL consecutively switches its state (turns ON or OFF) in a short period of time. |

conduction with the ambient, the second term $P^i\eta^i$ denotes the rate of energy transfer for the $i$-th TCL, and the last term $w^i$ accounts for external disturbances from occupancy, solar radiation, internal appliances, etc., and it is assumed to be drawn from a Gaussian distribution with zero mean [14], [17].

Each TCL has a temperature setpoint $\theta_s^i$ with a hysteretic ON/OFF local control within a deadband $[\underline{\theta}^i, \overline{\theta}^i]$ where $\underline{\theta}^i := \theta_s^i - \delta^i/2$ and $\overline{\theta}^i := \theta_s^i + \delta^i/2$. Table II lists the typical values for a residential Air Conditioning (AC) unit in summer time. (The parameter value are adapted from [17], [23].) The operating state $m^i(t)$ evolves as

$$m^i(t+\Delta t) = \begin{cases} 0 & \text{if } \theta^i(t+\Delta t) < \underline{\theta}^i, \\ 1 & \text{if } \theta^i(t+\Delta t) > \overline{\theta}^i, \\ m^i(t) & \text{otherwise,} \end{cases} \tag{2}$$

where $\Delta t \ll 1$ is a small time increment. The aggregated power consumption for all TCLs at time $t$ is given by

$$P(t) = \sum_{i=1}^{N} m^i(t) P^i. \tag{3}$$

### B. Discrete-Time Model

In simulation studies, we use the following discrete-time version of the above dynamics. Upon discretization, the temperature evolution dynamics (1) can be represented as a first-order difference equation

$$\theta_{k+1}^i = a^i \theta_k^i + (1-a^i)(\theta_a - m_k^i R^i P^i \eta^i) + \epsilon_k^i, \tag{4}$$

where $\theta_k^i$ is the internal temperature of the $i$-th TCL at time step $k$. The model parameter $a^i$ is defined as $a^i := e^{-h/C^iR^i}$, where $h$ is the discretization step. Moreover, the operating state $m_k^i$ can be written as

$$m_{k+1}^i = \begin{cases} 0 & \text{if } \theta_k^i < \underline{\theta}^i, \\ 1 & \text{if } \theta_k^i > \overline{\theta}^i, \\ m_k^i & \text{otherwise.} \end{cases} \tag{5}$$

Table II
TYPICAL PARAMETER VALUES FOR A RESIDENTIAL AC UNIT.

| Parameter | Description | Value | Unit |
|---|---|---|---|
| $C$ | thermal capacitance | 2 | kWh/°C |
| $R$ | thermal resistance | 2 | °C/kW |
| $P$ | rated electrical power | 5.6 | kW |
| $\eta$ | coefficient of performance | 2.5 | |
| $\theta_a$ | ambient temperature | 32 | °C |
| $\theta_s$ | temperature setpoint | 20 | °C |
| $\overline{\theta}$ | temperature upper bound | 21 | °C |
| $\underline{\theta}$ | temperature lower bound | 19 | °C |

The aggregated power consumption for all TCLs at time step $k$ is given by

$$P_k = \sum_{i=1}^{N} m_k^i P^i. \tag{6}$$

### C. Steady-State Operation

Consider a collection of heterogeneous TCLs whose thermal dynamics are described by Model (1)-(2). Under the assumption that the dynamics are noise free, we define $T_{\text{ON}}^i$ as the time it takes for the $i$-th TCL to transport from its upper temperature bound $\overline{\theta}^i$ to its lower temperature bound $\underline{\theta}^i$. From Model (1), it is straightforward to show that

$$T_{\text{ON}}^i := -R^iC^i \ \ln \frac{\underline{\theta}^i - \theta_a + R^iP^i\eta^i}{\overline{\theta}^i - \theta_a + R^iP^i\eta^i}. \tag{7}$$

Similarly, $T_{\text{OFF}}^i$ is defined as the time it takes for the $i$-th TCL to transport from its lower temperature bound $\underline{\theta}^i$ to its upper temperature bound $\overline{\theta}^i$. Formally,

$$T_{\text{OFF}}^i := -R^iC^i \ \ln \frac{\overline{\theta}^i - \theta_a}{\underline{\theta}^i - \theta_a}. \tag{8}$$

The duty cycle of the $i$-th TCL is defined as:

$$d_0^i := \frac{T_{\text{ON}}^i}{T_{\text{ON}}^i + T_{\text{OFF}}^i}. \tag{9}$$

At steady-state, the number of TCLs that are ON (or OFF) is a constant (see Fig. 2). The baseline power of all TCLs operating at steady state is given by

$$P_0 \equiv \sum_{i=1}^{N} P^i d_0^i. \tag{10}$$

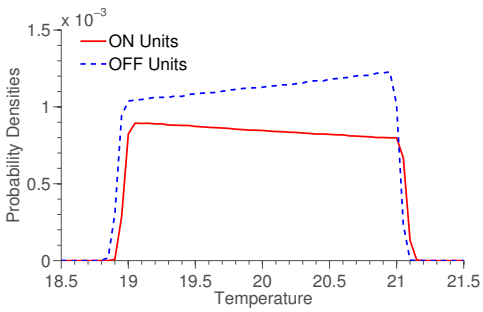



Figure 2. Approximate steady-state temperature distribution of a population of $10,000$ homogeneous TCLs. The used parameters are those listed in Table II with a discretization step of $h = 1$ and a noise level of $\text{var}(w) = 0.01$. The probability densities are calculated based on the number (normalized by $10,000$) of ON/OFF units in each temperature bin.

Fig. 2 depicts the approximate steady-state temperature distribution of a collection of TCLs. We observe that the temperature distribution curves are not flat. For example, there are fewer ON units at higher temperatures (close to the upper temperature bound) as compared to ON units whose temperatures are lower (close to the lower temperature bound). On the contrary, there are more OFF units at higher temperatures as compared to OFF units at lower temperatures. This phenomenon is due to the fact that the temperature of each unit changes slower when they are closer to the final temperature boundaries (i.e., the exponential rate of temperature change due to Model (1)).

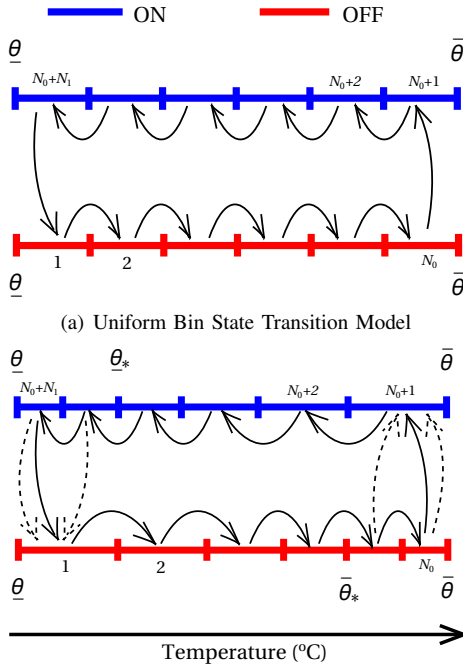
## III. Aggregate Modeling of a Population of TCLs

It is challenging to utilize individual hybrid state models in controlling a large population of TCLs. In this section, we develop an aggregate model to describe the temperature and operating state (ON or OFF) evolution of a collection of TCLs, which is amenable for control design. We propose a *non-uniform* bin state transition model which extends the uniform bin transition models considered in [19]–[21]. We first present our model for homogeneous TCLs and then discuss its extension to a heterogeneous case.

### A. Uniform Bin State Transition Model

In the uniform bin state transition model, the temperature deadband of each TCL is divided into *equally-spaced* bins (i.e., bins of the same length). Fig. 3 (a) depicts the uniform bin state model. Let $N_0$ be the number of bins in the OFF state and $N_1$ be the number of bins in the ON state. In a uniform model, $N_0 = N_1$. Each bin is associated with a *bin state*, resulting in a state vector whose dimension is equal to the total number of bins, $N_u = N_0 + N_1 = 2N_0 = 2N_1$. Each bin state represents the number of TCLs in a particular temperature bin with a particular state (ON or OFF).

While each temperature bin consists of a temperature range, in a uniform bin state transition model, a temperature




(a) Uniform Bin State Transition Model

(b) Non-uniform Bin State Transition Model

Figure 3. Uniform and Non-uniform bin state transition models. Each bin state represents the number of units whose temperatures lie within a particular temperature bin and with a particular state (ON or OFF). Solid arrows denote natural transitions of the units due to their local hysteretic control. Dashed arrows represent transitions due to manipulation of the units for ancillary service. Transitions in the non-uniform model are shown.

point (e.g., middle point) in that bin is selected as the representative bin temperature [21]. Based on this model, it is assumed that the population in each bin is transported to the next bin with a rate which is the inverse of the time it takes for a TCL to transport from the representative temperature of a bin to the representative temperature of the next level bin. While assigning one temperature (e.g., the middle point) to each bin as the representative temperature results in a simpler design of the state transition model, it ignores the temperature span of each bin and consequently ignores the fact that the units with different initial temperatures, even in one bin, have different transport times for reaching the next level bin. Moreover, because of the exponential behavior of the temperature change, the transition rate from one bin to the next level bin is different for each pair of neighboring bins. Even a more simplified model for temperature evolution of the TCLs is considered in [20] in which a uniform bin state model is considered while the transfer rates for ON or OFF states are the same and are related to the total evolution time of one TCL as its temperature changes from the set point temperature to the ambient temperature. As a consequence of these simplifications in the state-space model, a prediction error will be introduced at each sample time and will further propagate as the prediction steps increase. This observation motivates our derivation of the non-uniform bin state transition model.

### B. Non-uniform Bin State Transition Model

We propose a *non-uniform* bin state transition framework in which the temperature deadband $\delta$ of each TCL is divided into non-uniform bins (bins of different length). Fig. 3 (b) depicts an example of such a non-uniform bin model. This scheme gives us more flexibility in designing the bin structure as compared with the uniform bin models which assign equal length to each bin. For example, we can consider different number of bins for the OFF and ON states. The merit of this model is that it yields the same transition rate for each pair of neighboring bins. Moreover, with appropriate sampling time, it guarantees the transition of all units from one bin to the next bin. We consider a state-space representation of the underlying temperature evolutions of the OFF and ON units, where the state is defined by a vector $\boldsymbol{x} \in \mathbb{R}^{N_0+N_1}$, where $N_0$ and $N_1$ are the number of bins in the OFF and ON states, respectively. Without loss of generality, we number the bins as shown in Fig. 3. Let $N_{\text{OFF},1}(t)$ be the number of OFF units whose temperature lies in temperature bin 1 at time $t$. Similarly, $N_{\text{OFF},N_0}(t)$ be the number of OFF units whose temperature lies in temperature bin $N_0$, $N_{\text{ON},N_0+1}(t)$ be the number of ON units whose temperature lies in temperature bin $N_0+1$, and $N_{\text{ON},N_0+N_1}(t)$ be the number of ON units whose temperature lies in temperature bin $N_0+N_1$ at time $t$. We define the state vector $\boldsymbol{x}(t)$ as

$$\boldsymbol{x}(t) = \begin{bmatrix} N_{\text{OFF},1}(t) \\ N_{\text{OFF},2}(t) \\ \vdots \\ N_{\text{OFF},N_0}(t) \\ N_{\text{ON},N_0+1}(t) \\ N_{\text{ON},N_0+2}(t) \\ \vdots \\ N_{\text{ON},N_0+N_1}(t) \end{bmatrix} \in \mathbb{R}^{N_0+N_1}. \tag{11}$$

As mentioned earlier, we are particularly interested in how the state vector $\boldsymbol{x}$ changes over time. We define the state transition matrix $A$ such that

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}, \tag{12}$$

where $A$ is an $(N_0+N_1)\times(N_0+N_1)$ matrix whose $(i,j)$-th entry is the *rate* at which the units in bin $j$ are transported to bin $i$, where $j < i$. The diagonal entries of $A$ are such that $A$ has *zero column-sum*. We design the number and the length of bins in such a way that for each pair of neighboring bins, the time it takes for a unit to move from the left/right boundary of one bin to the left/right boundary of the next bin, $t_{\text{OFF}}$ or $t_{\text{ON}}$, is the same. Based on model (1) and considering our non-uniform bin model, we have

$$t_{\text{OFF}} = -\frac{RC}{N_0}\ln\frac{\overline{\theta}-\theta_a}{\underline{\theta}-\theta_a},\ t_{\text{ON}} = -\frac{RC}{N_1}\ln\frac{\underline{\theta}-\theta_a+PR\eta}{\overline{\theta}-\theta_a+PR\eta}.$$

Observe that in this setting the state matrix $A$ is a sparse matrix which has a structure as

$$A = \begin{bmatrix} -r_{\text{OFF}} & 0 & & \cdots & & 0 & +r_{\text{ON}} \\ +r_{\text{OFF}} & \ddots & 0 & & \cdots & & 0 \\ 0 & \ddots & -r_{\text{OFF}} & \ddots & & & \vdots \\ \vdots & \ddots & +r_{\text{OFF}} & -r_{\text{ON}} & & & \\ & \ddots & & +r_{\text{ON}} & -r_{\text{ON}} & & \\ & & & & \ddots & \ddots & 0 \\ 0 & \cdots & & & 0 & +r_{\text{ON}} & -r_{\text{ON}} \end{bmatrix},$$

where $r_{\text{OFF}} = \frac{1}{t_{\text{OFF}}}$ is the rate at which the bins in one bin are transported to the next bin when they are OFF, and similarly, $r_{\text{ON}} = \frac{1}{t_{\text{ON}}}$ is the rate at which the bins in one bin are transported to the next bin when they are ON. One should note that with a non-uniform bin transition framework, the rates at which the bins are transported in the OFF state are all equal. The same fact holds for bins in the ON state.

We next explain how to incorporate the manipulation (turn ON or OFF) of TCLs into the aggregate modeling scheme. Sate transition model (12) can be extended to

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x} + B\boldsymbol{u}, \tag{13}$$

where $\boldsymbol{u} \in \mathbb{R}^{N_0+N_1}$ is the control input and contains the number of ON or OFF units in each bin state considered for manipulation. One should observe that the control vector $\boldsymbol{u}$ has a similar structure as the state vector $\boldsymbol{x}(t)$. The control matrix $B$ contains the fractions ($|B_{i,j}| \leq 1$) of the units in one bin with a given state (OFF or ON) that are transported to the corresponding bin(s) in the other state. Fig. 3 shows such a manipulation scheme. As an illustrative example, the ON units in $(N_0+N_1-1)$-th and $(N_0+N_1)$-th bins are transported to the first bin when we perform an ON-to-OFF manipulation. Consequently,

$$B_{1,N_0+N_1-1} = B_{1,N_0+N_1} = +1,$$
$$B_{N_0+N_1-1,N_0+N_1-1} = B_{N_0+N_1,N_0+N_1} = -1.$$

Similarly, the OFF units in $(N_0-1)$-th and $N_0$-th bins are transported to the $(N_0+1)$-th bin when we perform an OFF-to-ON manipulation. Consequently,

$$B_{N_0+1,N_0-1} = B_{N_0+1,N_0} = +1,$$
$$B_{N_0-1,N_0-1} = B_{N_0,N_0} = -1.$$

The corresponding matrix $B$ in (13) incorporates such transitions when a manipulation of the units is performed.

From a practical point of view, the switching action of TCLs must respect operating constraints. For instance, when a TCL is turned OFF, it cannot be turned ON immediately. The unit has to be OFF for a certain amount of time to prevent short cycling. To address this issue, we introduce two thresholds, $\underline{\theta}^*$ and $\overline{\theta}_*$, where $\underline{\theta} < \underline{\theta}_*$, and $\overline{\theta}_* < \overline{\theta}$. We

impose the constraints that only the units with temperature in the interval $[\underline{\theta},\ \underline{\theta}_*]$ can be switched from ON to OFF. Similarly, only the units with temperature in the interval $[\overline{\theta}_*,\ \overline{\theta}]$ are allowed to switch from OFF to ON.

*Remark 3.1:* The control matrix $B$ can be built given the OFF and ON bin structure and the chosen thresholds ($\overline{\theta}_*$ and $\underline{\theta}_*$). One should observe that we only manipulate the units whose temperatures meet the considered thresholds. In the example shown in Fig. 3, the thresholds are chosen exactly equal to one of the bin boundaries for simplicity of illustration. However, a similar analysis can be done when the thresholds are chosen off the temperature grid. This results in a $B$ matrix whose some of its entries have magnitude smaller than one (i.e., $|B_{i,j}| < 1$). Such fractions (entries of $B$) can be found based on the considered non-uniform ON and OFF grids and how one bin in one state overlaps with the corresponding bins in the other state.

The aggregate power consumption of a homogeneous collection of TCLs at time $t$ is simply the sum of the power consumption of the units which are ON. Formally,

$$P_{\text{agg}}(t) = C\boldsymbol{x}(t),$$

where $C$ is the *measurement vector* whose first $N_0$ entries are zero and the rest $N_1$ entries are the rated power, $P$, as:

$$C = [\underbrace{0\ 0\ \cdots\ 0}_{N_0}\underbrace{P\ P\ \cdots\ P}_{N_1}] \in \mathbb{R}^{1\times(N_0+N_1)}.$$

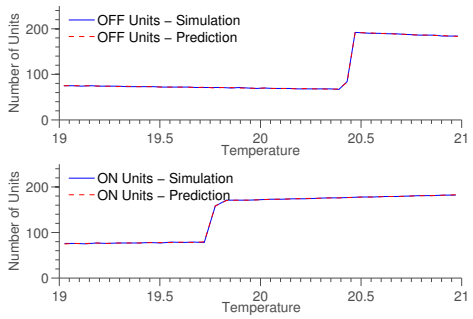
### C. Prediction Performance Comparison Between the Uniform and Non-uniform Bin State Models

In this section, we compare the prediction performance between our proposed non-uniform bin model and the uniform bin model. Recall that the benefit of this non-uniform bin model is that it guarantees, for a population of homogeneous TCLs subject to no noise, in each prediction horizon $\Delta t = t_{\text{ON}}$ (or $\Delta t = t_{\text{OFF}}$), the probability that all units in one bin (ON or OFF) transfer to the next level bin is 1. However, in a uniform bin model, the probability that all units in one bin move to the next level bin is not 1. As a result, at each prediction step, it introduces a prediction error, and this error will propagate as the number of predictions increases. In the simulation, we choose the prediction horizon as $\Delta t \approx (t_{\text{ON}} + t_{\text{OFF}})/2$ and $t_{\text{ON}} \approx t_{\text{OFF}}$. The state-space model is then discretized as

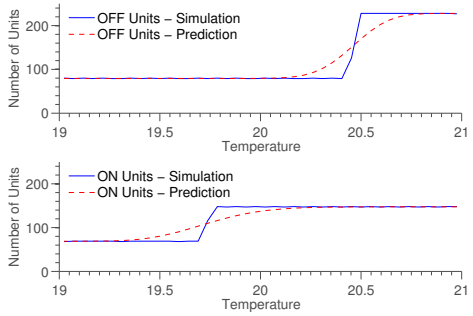
$$\boldsymbol{x}_{k+1} = A_k\boldsymbol{x}_k, \tag{14}$$

where the discrete-time state matrix $A_k$ is given by $A_k = (I_{N_0+N_1} + A)\Delta t$, where $A$ is the state transition matrix of the continuous-time model (12).

We first consider a population of homogeneous TCLs. The model parameters used in the simulation are as those given in Table II. We choose the number of ON and OFF state bins as $N_1 = 48, N_0 = 36$. As a result, the transfer rate is given by $r_{\text{OFF}} \approx r_{\text{ON}} \approx 1/50\ (s)$. For the uniform bin model,




(a) Non-uniform Bin Model

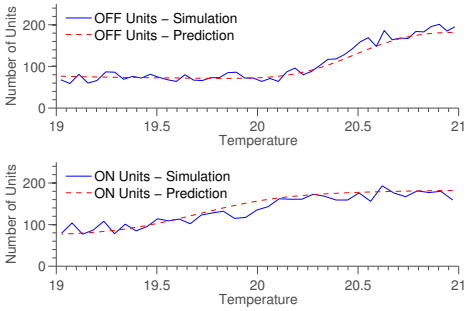



(b) Uniform Bin Model

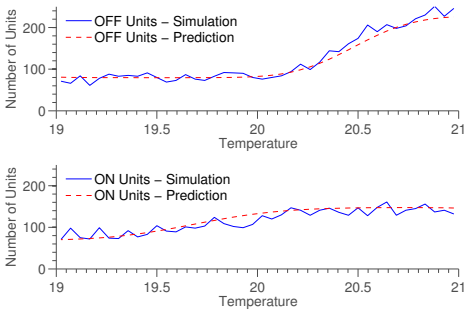
Figure 4. Prediction performance comparison between non-uniform and uniform bin models for a population of *homogeneous* TCLs with *no noise*.

we choose the number of bins as $N_u = N_1 + N_0 = 84$ (42 bins for both ON and OFF states). We choose $\Delta t$ in (14) as $\Delta t \approx 50\ (s)$, and simulate the population of TCLs for 60 minutes. We then compare the prediction performance of the two models at the last time step. Fig. 4 depicts the predictions for both non-uniform and uniform bin models. For the sake of comparison, we also include the true temperature distribution of the TCLs by simulating their temperature using Model (4)-(5) with discretization time step $h = 1\ (s)$. Based on the simulation results, the non-uniform model has a much better prediction than the uniform model for both the ON and OFF units; in each case its prediction error is less than 2 units. As long as the discretization step $\Delta t \approx t_{\text{ON}} \approx t_{\text{OFF}}$, the non-uniform model yields near-perfect prediction. Observe that the smaller number of bins, the larger of $\Delta t$ should be.

We next compare the prediction performance between the two models for a population of heterogeneous TCLs. The model parameters are obtained by introducing a $20\%$ heterogeneity to the typical values given in Table II. In addition, a Gaussian noise with zero mean and variance $0.01$ is added to the dynamics (1). In the simulation, we choose the discretization step $\Delta t = 25$ seconds. We observe from Fig. 5 that the non-uniform and uniform model have similar performance and they both make a good prediction of the temperature distribution of a collection of heterogeneous TCLs subject to noise. Moreover, we notice (by comparing




(a) Non-uniform Bin Model




(b) Uniform Bin Model

Figure 5. Prediction performance comparison between the non-uniform and uniform bin models for a population of *heterogeneous* TCLs with *noise*.

Fig. 4 vs. Fig. 5) that introducing heterogeneity and noise will damp the true temperature distribution of TCLs, making the distribution curves smoother. This is the reason behind choosing the discretization $\Delta t$ to be much smaller than $t_{\text{ON}}$ and $t_{\text{OFF}}$ so that the entries of $A_k$ in (14) are much smaller than 1, which means at each time step, only a small fraction of TCLs in that bin transport to the next level bin. This makes the non-uniform model presents a good prediction of the smooth distribution curves.

In summary, if the TCLs are homogeneous and noise free, then the non-uniform model with $\Delta t \approx t_{\text{OFF}} \approx t_{\text{ON}}$ makes a near perfect prediction. On the other hand, if heterogeneity and noise are introduced into the dynamics, then using a non-uniform model and a smaller discretization step result in a good prediction. For a population of TCLs with large heterogeneity, the non-uniform model can be applied by clustering the TCLs into several groups [21], in which each group has similar model parameters, and aggregately model each group by the proposed non-uniform bin model.

## IV. Decentralized Priority Control of TCLs for Regulation Service

So far we have presented a framework for aggregate modeling of a collection of TCLs. In this section, we propose a *decentralized priority control* strategy based on this aggregate model to control the power consumption of TCLs and in order to provide fast regulation service to the grid. Fig. 6 illustrates the proposed control architecture. The

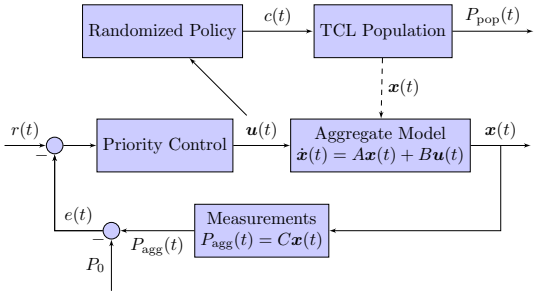



Figure 6. The proposed decentralized priority control architecture for provision of fast regulation service.

control design is based on the aggregate model of TCLs. At each sample time, a control signal $c(t)$ is constructed based on $\boldsymbol{u}(t)$ (the control input) from the aggregate model. The control signal $c(t)$ contains a switching probability based on which a unit in a temperature bin turns ON or OFF. This signal is then broadcast to the population of TCLs. One should note that the control strategy is decentralized in the sense that each TCL makes its control decision (turning ON or OFF) solely based on this *common* control signal $c(t)$. At each sample time, each TCL generates a random number drawn from an uniform distribution between 0 and 1 and compares this number with the corresponding switching probability $c(t)$ that has received from the controller.

The regulation signal sent by the system operator is typically a sequence of pulses at 4 second intervals [24]. In the case of loads engaged in regulation provision, the magnitude of the pulse is the required amount of power deviation from their baseline asked by the grid operator. Suppose that the population of TCLs is required to provide $r(t)$ (in MW) amount of regulation service at time $t$. The control objective is to manipulate the power consumption of TCLs so that the difference between $P_{\text{agg}}(t)$, the power consumption of the collection predicted by the aggregate model, and the baseline power $P_0$, $e(t) = P_{\text{agg}}(t) - P_0$, tracks this regulation signal $r(t)$. One should note that the ultimate goal of the controller is to manipulate the collection of TCLs such that the actual power consumption of TCLs, denoted by $P_{\text{pop}}(t)$ in Fig. 6, follows the regulation signal. In order to increase the accuracy of the aggregate model, we update the state-space model with our measurements from the population. The measurements needed for implementing such a state update is the temperature and the ON/OFF state of each TCL. Based on such information, we can construct the state vector, $\boldsymbol{x}(t)$. Consequently, $\boldsymbol{x}(t)$ can be used to update the state estimate in the aggregate model. Observe that we use a dashed line to illustrate this update to show that such an update can be performed whenever needed and to increase the accuracy of the aggregate model. In other words, this update is not necessarily needed at every sample time and based on the dynamics of the system under study can be performed at larger time intervals.

As an example, when TCLs are supposed to consume

less power (i.e. $r(t) < 0$) in order to provide energy to the grid, the controller turns OFF a particular number of ON units. A similar action will be taken when TCLs are required to consume more power. In this paper, we propose a decentralized priority control strategy in which we assign higher switching (respectively, turn ON or OFF) priority to the units whose temperature are close to the (respectively, upper or lower) temperature bound.

Specifically, when OFF units are being manipulated, we assign the highest switching priority to the units in the Bin $N_0$ and progressively assign less and less priority to the units in the left-hand side bins (see Fig. 3). Note that the units in the same bin have the same priority. Similarly, when ON units are being considered, we assign the highest priority to the units in Bin $N_0 + N_1$ and gradually assign less and less priority for the units in the right-hand side bins. Based on the regulation procurement $r(t)$, the priority control strategy first calculates the required number of bins with higher priorities so that the manipulable power deviation of TCLs in these bins is larger than or equal to the regulation $r(t)$. In practice, when the manipulable power deviation is larger than $r(t)$, the needed number of units in the bin with the least priority among found is more than enough. Hence, we do not need to manipulate all of them in that bin. To this end, we use a randomized policy. Suppose, we only need to manipulate $N_n$ units of all units in the bin with the least priority. Also assume that at time $t$ there are $N_a$ available units in that particular bin. In order to manipulate those $N_n$ units, we generate a switching probability with value $N_n/N_a$. Observe that the switching probability for the units who lie in bins of higher priority is 1. Formally, the switching probability, when units are required to *consume more* power (i.e., turning ON some of the OFF units), can be expressed as:

$$c^{\text{ON}}(t) = \begin{cases} N_n/N_a, & \text{if } \theta^- \le \theta^i(t) < \theta^+, \\ 1, & \text{if } \theta^+ \le \theta^i(t), \end{cases} \tag{15}$$

where $\theta^-$ and $\theta^+$ are respectively the lower and upper temperature bounds for the bin with the least priority. Similarly, the switching probability, when units are required to *consume less* power (i.e., turning OFF some of the ON units), can be expressed as:

$$c^{\text{OFF}}(t) = \begin{cases} N_n/N_a, & \text{if } \theta^- \le \theta^i(t) < \theta^+, \\ 1, & \text{if } \theta^i(t) \le \theta^-, \end{cases} \tag{16}$$

where $\theta^-$ and $\theta^+$ are respectively the lower and upper temperature bounds for the bin with the least priority. The controller then broadcasts this probability signal $c(t)$ to the population of TCLs. Locally, each unit generates a random number, compares it with the received probability signal, and takes a suitable switching action. Such a randomized decentralized control strategy reduces the communication requirement for control implementation. The priority control algorithm is summarized in Algorithm 1.

**Algorithm 1** Decentralized Priority Control Algorithm

**loop**
  obtain $r(t)$;
  define $e(t) = P_{\text{agg}}(t) - P_0$;
  **if** $e(t) < r(t)$ **then**
    find $i^* = \max\left\{i \mid \sum_{j=i}^{N_0} \boldsymbol{x}_j \ge \frac{r(t)-e(t)}{P}\right\}$;
    define $N_n := \sum_{j=i^*}^{N_0} \boldsymbol{x}_j - \frac{r(t)-e(t)}{P}$;
    define $N_a := \boldsymbol{x}_{i^*}$;
    construct $c^{\text{ON}}(t)$ using (15);
    broadcast $c^{\text{ON}}(t)$;
  **else if** $e(t) > r(t)$ **then**
    find $i^* = \max\left\{i \mid \sum_{j=i}^{N_0+N_1} \boldsymbol{x}_j \ge \frac{r(t)-e(t)}{P}\right\}$;
    define $N_n := \sum_{j=i^*}^{N_0+N_1} \boldsymbol{x}_j - \frac{e(t)-r(t)}{P}$;
    define $N_a := \boldsymbol{x}_{i^*}$;
    construct $c^{\text{OFF}}(t)$ using (16);
    broadcast $c^{\text{OFF}}(t)$;
  **end if**
**end loop**

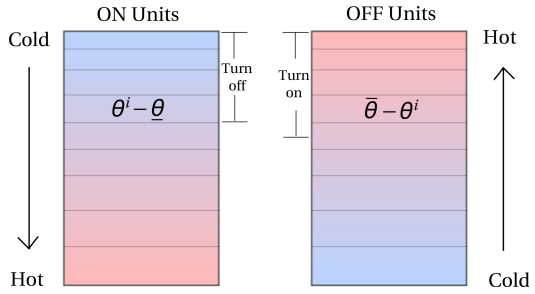



Figure 7. A pictorial representation of priority stacks.

## V. Numerical Experiments

In this section, we apply our proposed modeling and control schemes to follow a regulation signal $r(t)$ taken from a randomly chosen 1-hour long sample from Pennsylvania-New Jersey-Maryland Interconnection (PJM) [24]. The magnitude of the original signal is scaled appropriately to match the power capacity of $10,000$ residential AC units with parameters as given in Table II.

The regulation signal is an AGC signal that contains system operator's setpoint commands. The AGC signal is built based on the data from the Supervisory Control and Data Acquisition (SCADA) system and the system frequency at every $4$ seconds. It is constructed in order to maintain the power system's frequency at a desired level by balancing control area's generation and load. There exists some performance compliance statistics that quantify how well an AGC signal is successful in balancing generation and load and in frequency control. Analyzing such statistics is out of the scopes of this paper.

### A. Manipulation of TCLs With Short Cycling

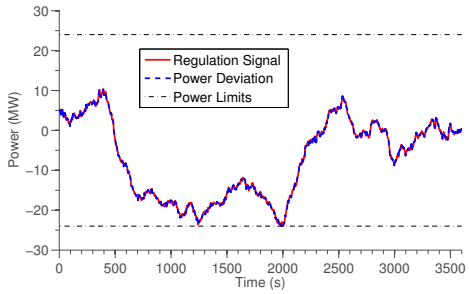
We first consider the case when there is no threshold constraint in the manipulation, which means short cycling

is allowed (e.g., for some types of TCLs such as water heaters). The parameters of each TCL is given in Table II. Based on (10), the baseline power of a population of $10,000$ homogeneous TCLs is $24$ MW and their aggregate rated power (when all units are in the ON state) is $56$ MW. In this scenario, we see from Fig. 8 (a) that when the magnitude of the downward regulation signal is less than $24$ MW, the power deviation, which is defined as the difference between the actual power consumption of TCLs, $P_{\text{pop}(t)}$, and the baseline power consumption given in (10), tracks the regulation signal very well. Additional simulations (not reported here) show that the power deviation of TCLs also tracks the regulation signal very well when their upward magnitude is smaller or equal to 24 MW. This shows that TCLs are capable to ramp up and down to their maximum available capacity (around $40\% - 45\%$ of the rated power) in time scales of a few minutes, demonstrating a large potential for TCLs in providing fast regulation service.

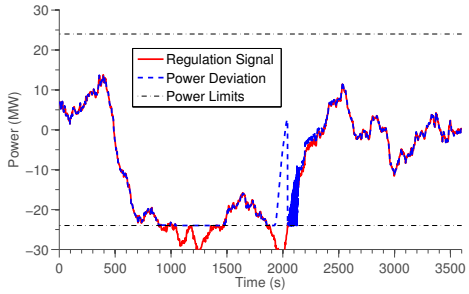
However, it is shown in Fig. 8 (b) that if the maximum magnitude of the downward regulation signal is larger than the baseline power, it results in a poor power tracking performance. More specifically, at about time $t = 900$ seconds the power deviation remains at $-24$ MW. This is because at $t = 900$ seconds all units are turned OFF, while their temperatures are still lower than the upper bound. In such conditions, all units remain OFF. However, another large downward regulation signal happens around time $t = 1900$ seconds. In contrast to what happened at $t = 900$ seconds, the TCL population can not maintain the $-24$ MW power deviation at $t = 1900$ seconds. This is because at this time some of the OFF units have reached the temperature upper bound and consequently, their local hysteretic controllers make them turn ON. However, they immediately turn OFF again due to a large downward regulation signal. The observed oscillation of the power deviation at $t = 2100$ seconds in Fig. 8 (b) is a result of this frequent ON/OFF switching of the units around the temperature upper bound. A similar situation can be observed when the upward regulation is larger than $56 - 24 = 32$ MW.

### B. Manipulation of TCLs Without Short Cycling

In this section, in order to avoid short cycling we introduce temperature thresholds in the manipulation scheme. These thresholds keep a unit from turning ON and OFF too frequently. In other words, we add a constraint to the manipulation process such that when a unit changes its state (either it turns ON from an OFF state or it turns OFF from an ON state), it should remain in that state for a certain amount of time that is suggested by the manufacturer and in order to avoid short cycling. In the simulation, we take this time equal to 10 minutes. Fig. 9 illustrates the tracking performance where units are avoided from short cycling. The result in Fig. 9 (a) indicates that with the same regulation signal as in Fig. 8 (a), the population of TCLs




(a) Tracking a regulation signal with moderate magnitude.




(b) Tracking a regulation signal with large magnitude.

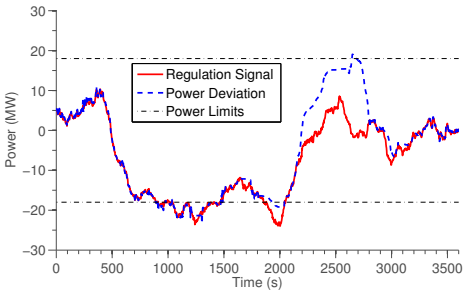
Figure 8. Performance of the proposed aggregate modeling and decentralized control schemes in tracking a regulation signal by manipulating a population of $10,000$ TCLs when short cycling of the units is allowed.

cannot track the regulation signal very well. This is because introducing the threshold constraints decreases the power capacity of the collection of TCLs. In Fig. 9 (b), we reduce the magnitude of the downward regulation signal to $18$ MW. In this case, the population of TCLs can accurately track the regulation signal. A similar performance is observed when the magnitude of the upward regulation is less than or equal to $18$ MW. This shows that even with threshold constraint, a population of TCLs can provide at least $30\%$ of their rated power for upward and downward regulation, without causing short cycling.

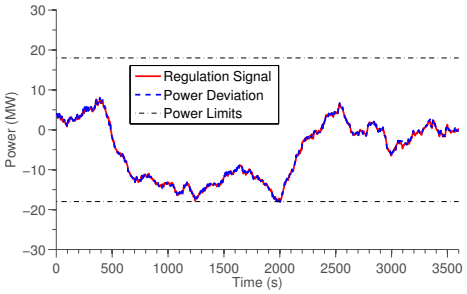
## VI. Conclusions and Future Work

In this paper, we first considered aggregate modeling of a population of TCLs where we proposed a non-uniform bin state transition model. The proposed model yields better prediction performance and requires fewer number of bins. We then proposed a randomized priority control strategy for manipulating a collection of TCLs for fast regulation service. The proposed control scheme is decentralized and results in reduced communication and implementation requirements.

Studying different ways of incentivizing users for participation is one of the issues that is under our current consideration. Among others, lottery-based methods seems to be promising as users usually prefer to participate in programs that have a large reward (although with very low probability) as compared to programs that promise small rewards with much higher probabilities. A systematic characterization of




(a) Tracking a regulation signal with large magnitude.




(b) Tracking a regulation signal with moderate magnitude.

Figure 9. Performance of the proposed aggregate modeling and decentralized control schemes in tracking a regulation signal by manipulating a population of $10,000$ TCLs when short cycling of the units is not allowed.

the power and energy capacity for a collection of heterogeneous TCLs as a function of ambient temperature and external disturbances from occupancy is another direction of research. We are also examining other sources of flexibility such as residential pool pumps, washing and dryer machines. Characterizing flexibility and scheduling of these deferrable loads for ancillary service is under current investigation.

## REFERENCES


[1] B. Kirby, "Ancillary services: Technical and commercial insights," Tech. Rep., July 2007. [Online]. Available: http://www.consultkirby.com/files/Ancillary_Services_-_Technical_And_Commercial_Insights_EXT_.pdf

[2] "Storage participation in ERCOT (prepared by the Texas energy storage alliance)," January 2010. [Online]. Available: http://www.ercot.com/

[3] 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, August 2007.

[4] 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.

[5] S. Meyn, M. Negrete-Pincetic, G. Wang, A. Kowli, and E. Shafieepoorfard, "The value of volatile resources in electricity markets," in *CDC2010*, 2010, pp. 1029–1036, and submitted to IEEE TAC, 2012.

[6] 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.

[7] Market and Infrastructure Policy, "2013 flexible capacity procurement requirement," Tech. Rep., March 2012. [Online]. Available: http://www.CAISO.com/

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

[9] 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, June 2008.

[10] MISO, "Frequency regulation compensation - FERC order no. 755," Tech. Rep., March 2013. [Online]. Available: https://www.midwestiso.org

[11] H. Hao, A. Kowli, Y. Lin, P. Barooah, and S. Meyn, "Ancillary service for the grid via control of commercial building HVAC systems," in *American Control Conference*, June 2013.

[12] F. Schweppe, R. Tabors, J. Kirtley, H. Outhred, F. Pickel, and A. Cox, "Homeostatic utility control," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-99, no. 3, pp. 1151–1163, May 1980.

[13] S. Ihara and F. C. Schweppe, "Physically based modeling of cold load pickup," *IEEE Transactions on Power Apparatus and Systems*, no. 9, pp. 4142–4150, 1981.

[14] 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.

[15] PG&E, "Smart AC program." [Online]. Available: http://www.pge.com/en/myhome/saveenergymoney/energysavingprograms/smartac/index.page

[16] FPL, "On call savings program." [Online]. Available: http://www.fpl.com/residential/energy_saving/programs/oncall.shtml

[17] 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.

[18] 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.

[19] S. Koch, J. Mathieu, and D. Callaway, "Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services," in *Proc. PSCC*, pp. 1–7, 2011.

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

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

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

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

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