

Load-Level Control Design for Demand Dispatch With Heterogeneous Flexible Loads

Joel Mathias^{ID}, Ana Bušić^{ID}, *Member, IEEE*, and Sean Meyn^{ID}, *Fellow, IEEE*

Abstract—Over the past decade, there has been significant progress on the science of load control for the creation of *virtual energy storage*. This is an alternative to demand response, and it is termed *demand dispatch*. Distributed control is used to manage millions of flexible loads to modify the power consumption of the aggregation, which can be ramped up and down, just like discharging and charging a battery. A challenge with distributed control is heterogeneity of the population of loads, which complicates control at the aggregate level. It is shown in this article that additional control at each load in the population can result in a far aggregate model. The local control is designed to flatten resonances and produce approximately all-pass response. Analysis is based on mean-field control for the heterogeneous population; the mean-field model is only justified because of the additional local control introduced in this article. Theory and simulations indicate that the resulting input–output dynamics of the aggregate has a nearly flat input–output response: the behavior of an ideal, multi-GW battery system.

Index Terms—Controlled Markov processes, demand dispatch, load frequency control, smart grids.

I. INTRODUCTION

IT IS clear to any user of electricity that many loads are flexible, and for this reason, there is consensus that load flexibility has great potential for the creation of virtual energy storage (VES). Consider, for example, the collection of all residential refrigerators in Florida. The consumer is usually concerned with a single quality of service (QoS) metric: that food is kept at the right temperature. So long as this constraint is satisfied, there is flexibility in terms of shifting the power consumption.

There are approximately 10 million homes in Florida. If each has a 200-W refrigerator, with 100-W average power consumption, then the nominal power consumption is 100×10^7 W, which is 1 GW.

Manuscript received 6 September 2022; accepted 11 February 2023. Date of publication 2 March 2023; date of current version 22 June 2023. This work was supported in part by the National Science Foundation under Grant EPCN-1935389. Recommended by Associate Editor N. Quijano. (Corresponding author: Joel Mathias.)

Joel Mathias is with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: joel.mathias@asu.edu).

Ana Bušić is with INRIA, 75012 Paris, France, and also with the Département d'Informatique de l'ENS, 75012 Paris, France (e-mail: ana.busic@inria.fr).

Sean Meyn is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: meyn@ece.ufl.edu).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCST.2023.3245287>.

Digital Object Identifier 10.1109/TCST.2023.3245287

As a striking, if absurd, thought experiment, suppose that a grid operator wishes power deviation (from nominal) to track a square wave with a 1-h time period. A control solution is formulated to create synchronization: each refrigerator turns off for 30 min and then on for 30 min in synchronization with other refrigerators, and the cycle repeats. This will ensure that the power deviation of the collection of refrigerators tracks a ± 1 -GW square wave with a 1-h time period. For a homogeneous population whose nominal duty cycle matches this signal, the internal temperature will remain within preset bounds, and no resident will realize that their refrigerator is providing a service to the grid.

This is one example of a simple task for one homogeneous population of loads (ignoring the impact of usage, door opening and closing, etc.). With the right mixture of local and global control, it is possible to obtain battery-like services for the grid, while respecting the QoS constraints of the loads and taking into account model imperfections and usage.

II. WHAT IS DEMAND DISPATCH?

The term demand dispatch was first introduced by Brooks et al. [7]. The following words are taken from the technical report written for the department of energy (DoE) [18]: demand dispatch is “an operating model used by grid operators to dispatch “behind-the-meter” resources in both directions—increasing and decreasing load as viewed at the system level—as a complement to supply (generation) dispatch to more effectively optimize grid operations.”

This differentiates it from the traditional demand response defined by FERC as “changes in electric use by demand-side resources from their normal consumption patterns in response to changes in the price of electricity, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized” [37].

In more concrete terms, demand dispatch refers to the control of flexible loads for the creation of VES—see [14], [19], [22] and the references therein. Automation is central to the success of demand dispatch: through intelligent control design, flexible loads, such as electric water heaters, refrigerators, and air conditioners, provide grid (supply–demand) balancing and regulation services while also ensuring consumer-side QoS.

A. Demand Dispatch Architecture

The following goals are the basis of the control architecture considered in this article: 1) distributed control to simplify

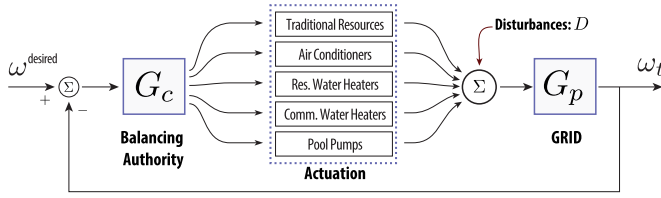


Fig. 1. Macro control architecture.

communication requirements and assuage consumer privacy concerns; 2) reliable ancillary service; and 3) strict bounds on consumer QoS. The final two goals are made possible through appropriate control at the load.

This article is not concerned with how incentives will be created to ensure participation, but there are plenty of examples today. Consider, for example, the use of contractual agreements and periodic credits, such as those proffered by Florida power and light (FPL) in their OnCall program and Austin Energy's Power Partner program that successfully incentivize customer participation.

Fig. 1 shows a macro control system diagram for regulating the grid frequency ω_t (at time t), with ω^{desired} being the ideal operating frequency of the grid (60 Hz in the U.S.), in the presence of disturbances. The disturbances can be due to forecast errors, high-frequency changes in renewable generation, generator trips, transmission line failures, demand surges, etc.

The compensation block with the associated transfer function G_c represents today's balancing authority (BA), and the "grid" represents the aggregate dynamics of loads, generators, transmission lines, and other grid elements. Design of G_c will be based on available resources (labeled *actuation* in the figure) and the grid transfer function G_p . This part of the control architecture is essentially what is used in most regions of the world today [27], [42].

Actuation includes flexible loads participating in demand dispatch along with "traditional resources" such as spinning reserves, fast-start generators, and batteries. It is argued here and in prior work that flexible loads can provide balancing services at a far lower cost than utility scale batteries [2], [11], [34], [40].

We have left out regulation of tie-line error, focusing on ramp services, balancing reserves, and frequency regulation. Incorporation of tie-line regulation is avoided only for simplicity of exposition.

B. The Heterogeneity Challenge

Returning once more to the example of residential refrigerators, it is unlikely that each has a duty cycle of period precisely 1 h. Moreover, there is usage, so that some percentage of refrigerators will require more power than nominal at any period in the day. It is clear from [23] that heterogeneity presents complications in aggregate modeling and capacity estimation. While Hao et al. [23] provide bounds on capacity for a heterogeneous population, the bounds are far more accurate under homogeneity.

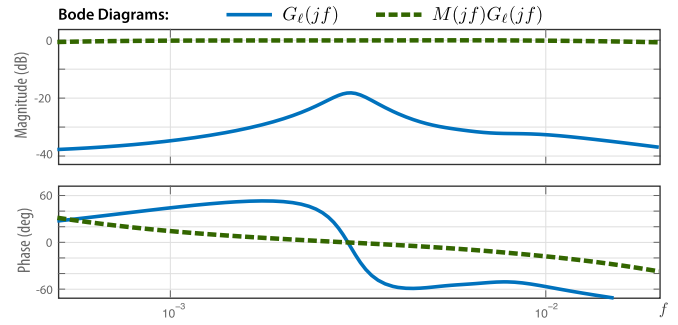


Fig. 2. Linearized mean-field dynamics: with and without inverse filter.

There is no inherent reason why heterogeneity should present major challenges for control. Consider, once more, the tracking problem involving a population of residential refrigerators. Suppose that the duty cycles are all 50%, but the periods vary from 15 to 90 min. Perfect tracking of a square wave remains possible (subject to some capacity loss due to usage). The only cost to consumers is that cycling may be increased or decreased from nominal. If this is a concern, then techniques from [17] can be used to enforce constraints on cycling, subject to a quantifiable loss in capacity.

The question then is how to formulate a model for control when the population is heterogeneous. The approach in this article is to design additional control at the load so that the aggregate model looks simple from the point of view of BA.

The general idea is formulated as follows. By design, if we did have a homogeneous population, then we can apply standard mean-field theory to obtain a dynamical system model. A time t , the input is denoted ζ_t , and the output y_t equals the average power consumption over the population. The goal of the BA is to design ζ based on available measurements so that y tracks a desired reference signal. Examples of a reference signal include the balancing reserves used in the Pacific Northwest, or the area control error (ACE) signal used in the majority of BAs.

While nonlinear by design, the dynamics are assumed smooth, and hence, they may be linearized about nominal behavior. We opt for a transfer function description of the linearization, denoted G_ℓ . Fig. 2 shows an example for a homogeneous collection of TCLs (in this example, small residential air conditioners). The resonance observed at frequency $f_r = 3 \times 10^{-3}$ rad/s is consistent with a cycling period of approximately 30 min. See Section IV for details on the construction of the transfer functions plotted in the figure.

The transfer function M appearing in the figure is an approximate inverse of G_ℓ , obtained using a robust control technique developed in the dissertation [30]. The result is essentially a notch filter that effectively suppresses the resonance. The BA broadcasts the signal U , but the load uses the transformed signal $\zeta = MU$ in the Laplace domain, and in the time domain the convolution

$$\zeta_t = \int_0^t h_{t-\tau} U_\tau d\tau, \quad t \geq 0 \quad (1)$$

also written $\zeta = h * U$, with h the impulse response associated with M . By design, the linearized aggregate dynamics are flat within a large range of frequencies.

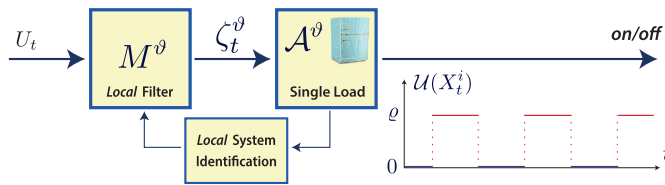


Fig. 3. Layers of local control at an individual load of type ϑ : 1) power transitions depend on the signal ζ_t^ϑ and the rate matrix A^ϑ ; 2) ζ_t^ϑ locally generated based on the grid-level signal U_t and the local filter M^ϑ ; and 3) System identification at the load to tune parameters defining M .

This argument holds even for a heterogeneous population, provided the filter M is specified at the load. It is assumed that the essential characteristics of the load are summarized by a vector $\vartheta \in \mathbb{R}^m$, which determines the transfer function G_ℓ^ϑ for a hypothetical linear model that would be obtained from a homogeneous population of identical loads with this parameter. Fig. 3 is introduced to highlight notation and emphasize that the control solutions investigated in this article are largely local: the filter M^ϑ with impulse response h^ϑ may be unique to a single load in the population, along with the rate matrix A^ϑ that together with $\zeta^\vartheta = h^\vartheta * U$ determines local on/off decisions. Software at an individual load estimates ϑ to create and/or update A^ϑ and M^ϑ based on data collected locally.

Note that in practice it may be more efficient to send different signals to different classes of resources (in Fig. 1 there are five classes). For example, PJM filters the AGC signal to generate RegA and RegD signals, the latter being sent to faster responding resources. Further discussion may be found in Section V-A.

C. Contributions

The main contributions of this article are summarized in the following.

- 1) Demand dispatch models are constructed for a *heterogeneous* population of loads operating in continuous time. Proposition 1 and its corollary concern mean-field limits and input–output response for a heterogeneous collection of loads.
- 2) Approaches to filter design are proposed based on techniques from the theory of robust control. This step is facilitated by a minimum phase property that is commonly observed in the linearized mean-field dynamics.
- 3) A survey of results from numerical experiments is provided. It is found in an ideal setting that tracking is nearly perfect with only one-way communication from the BA to the loads. With the introduction of unmodeled dynamics (such as 50% swings in capacity), there is some performance degradation, but grid-level frequency remains within required bounds. This is possible because of global feedback (see Fig. 1).

A striking finding in Section VI is the impact of the daily periodic patterns of response from loads such as air conditioning. In simulations, it is found that the grid frequency is maintained within tight bounds, even though the capacity varies by $\pm 50\%$.

A version of the filter design was proposed without analysis in the conference proceedings [31], and versions of the main

technical results Propositions 1 and 2 may be found in the first author's dissertation [30].

We are not aware of any prior work providing mean-field theory for a population of heterogeneous loads; prior results were restricted to homogeneous load aggregations, or heuristics. An example of a heuristic is to create a homogeneous model for control by averaging over the population coefficients of local ODE models [13], [34].

D. Related Research

An aggregate model for a large aggregation of flexible loads was first proposed in [29], expressed as a stochastic differential equation. Much later came more tractable mean-field models. In particular, Koch et al. [26], [34] introduced randomized algorithms and mean-field models for control.

In both [26], [34] and the later work [40], bilinearity of a mean-field model is exploited for the purposes of feedback linearization (this bilinearity can be anticipated from (13) below on relaxing any structure imposed on the controlled differential generator). These prior works also considered the impact of heterogeneity: in [26], [34], and later [13], the linear model parameters were simply averaged for the purposes of control synthesis, and the resulting control solution was tested through simulation. [40, Sec. 5] contains numerical results that suggest that their method is well-suited to tracking even for a heterogeneous population. Their approach amounts to a form of dead-beat control (which raises questions regarding robustness), but this approach is worthy of further research.

The work of [26], [34], and [40] does not fit well into the framework of this article, since this approach requires state estimation for the mean-field model which presents its own challenges [13]. The control approach of [23] for TCLs also falls out of scope, since it requires full information for each load so that a priority approach can be used to determine which loads the BA turns on or off at each time.

These challenges motivated the approach of [36] and [35] based on local randomized policies. This initial work focused on residential pool pumps, but the same control architecture can be used for other on/off loads [8], [14]. Similar strategies were pursued in [3] and [2], whose algorithms might benefit from the filtering approaches proposed here. Another approach to dealing with heterogeneity is discussed in [20]: different command signals are generated for different load classes based on a solution of a quadratic program. The optimization problem is solved by a centralized coordinator, which differentiates it from the local control of this article.

Camardella et al. [9] and Chertkov and Chernyak [15] contain methods and history of related approaches that focus on finite-horizon stochastic optimal control (which is likely an ideal setting for peak shaving and ramping services).

The focus of the simulations in this article is on ramping and AGC. We do not consider primary frequency response. However, a discussion on the randomized control of flexible loads for primary frequency response can be found in [4], [41], and [43]. A study of the robustness of the control design is required for the participation of loads in primary frequency response [32].

Capacity bounds for VES were introduced in [23] for TCLs and [35], [36] for residential pool cleaning (which also applies to similar loads such as water pumping for sanitation or irrigation). Cycling constraints are accounted for in [17], [38], [44], and [16].

Organization: The remainder of this article is organized as follows. Section III provides details of the distributed control architecture. Filter design to create approximate homogeneity is the focus of Section IV, which concludes with analysis based on the (infinite dimensional) mean-field model for a heterogeneous population.

The results from numerical experiments are summarized in two sections: Section V surveys the results from simulations of a heterogeneous population of Markovian loads, with comparisons to the mean-field model and its linearization. Section VI summarizes grid-level experiments to demonstrate the validity and utility of demand dispatch. Conclusions and directions for future research are contained in Section VII.

III. DISTRIBUTED CONTROL ARCHITECTURE

The following conventions are adopted for state, control, and output variables: uppercase letters are used to denote random variables, and boldface is used to denote functions of time; for example, x_t is a deterministic state at time t , X_t is a random variable, and $\mathbf{X}$ is a random process.

A. Grid-Level Control

The macro grid model used in this study is a linear input–output model in continuous time in which the input is power deviation, and the output is frequency deviation, following standard practice [27]. A particular example from [12] is used in numerical experiments

$$G_p(s) = 10^{-5} \frac{2.488s + 2.057}{s^2 + 0.3827s + 0.1071}. \quad (2)$$

The impulse response of this system is in close agreement with the response of frequency to a grid outage in the ERCOT region—a full discussion can be found in [12].

The resonance of this transfer function corresponds to time-scales on the order of seconds, while control problems of interest in this article concern disturbance rejection on much slower time-scales. The flat dynamics at low frequencies is the reason why BAs use PI control to synthesize AGC.

The same approach was used in the numerical experiments surveyed below, so that $G_c(s) = K_P + K_I/s$ in Fig. 1. The following was adopted in the experiment surveyed in Section VI, where the choice of parameters was based on the understanding that there is significant model uncertainty at high frequencies

$$G_c(s) = 516 \frac{s + 0.5}{s}. \quad (3)$$

Fig. 4 shows the Bode plots of G_p , G_c , and the loop transfer function $G_p G_c$.

The time-domain description of this PI compensator is expressed as

$$U_t = K_P \tilde{\omega}_t + K_I \int_0^t \tilde{\omega}_r dr, \quad \tilde{\omega}_t = \omega^{\text{desired}} - \omega_t.$$

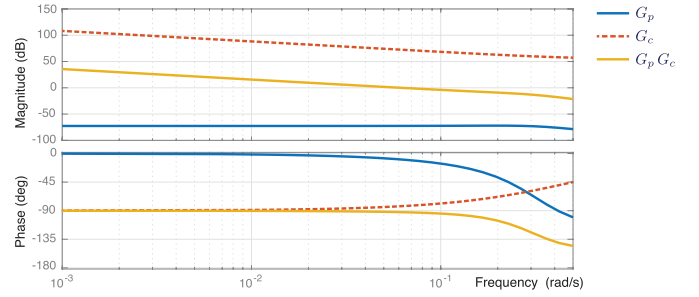


Fig. 4. Bode plot of the grid G_p , the PI compensator G_c , and the open-loop transfer function $G_p G_c$. The 0-dB crossover frequency of the open-loop transfer function is 0.06 rad/s, and the corresponding phase margin is approximately 90deg.

The signal U_t can be interpreted as the desired change in power (MW) from all the resources. The control gains K_P and K_I are chosen to respect the uncertainty of grid dynamics on time-scales of seconds or faster.

Remember that U_t is a command signal, and physical resources are required for actuation. Fig. 3 illustrates how U_t is received and transformed at an individual load in a demand dispatch architecture. The details are explained next.

B. Local Control: Markovian Load Dynamics

The behavior of a TCL or residential pool pump is largely deterministic. For example, a pool pump has a fixed duty cycle, and a residential refrigerator is approximately periodic with a period of about 1 h, with 50% duty cycle. Disturbances for TCLs come in the form of external temperature variations and usage [14].

A Markov model is adopted so that we can smoothly modify power consumption, and in particular justify a linearization of the corresponding mean-field model. Justification will be clearer after we provide details.

1) *Nominal Model and Perturbations:* The state process evolves in continuous time, on a finite state space with d elements, denoted $\mathbf{X} = \{x^1, \dots, x^d\}$. The nominal model is a Markov process, whose dynamics are defined by a *rate matrix* $\mathcal{A}_0$. The transition semi-group is defined by the exponential family: for any $t \geq 0$ and $x, x' \in \mathbf{X}$

$$P^t(x, x') := \mathbb{P}\{X_t = x' | X_0 = x\} = \exp(t \mathcal{A}_0(x, x')). \quad (4)$$

It is assumed that $\mathbf{X}$ is ergodic, i.e., the nominal model has a unique invariant probability mass function (pmf), denoted π_0 . Invariance is equivalent to the identity $\sum_x \pi_0(x) \mathcal{A}_0(x, x') = 0$ for every $x' \in \mathbf{X}$.

It is assumed that the state space has the product form $\mathbf{X} = \mathbf{X}^u \times \mathbf{X}^n$, and we write $X_t = (X_t^u, X_t^n)$ for $t \geq 0$. The first component $\mathbf{X}^u$ represents a variable that can be adjusted directly, which will be called *nurture*. We focus on on/off loads in this article, so that $\mathbf{X}^u = \{0, 1\}$. The second component $\mathbf{X}^n$ is called *nature*, and it is only *indirectly* influenced by $\mathbf{X}^u$ and exogenous disturbances; for example, the nature component may coincide with the internal temperature of a TCL.

While not essential, it is convenient to restrict to the following structure:

$$\mathcal{A}_0 = r[-I + S_0] \quad (5)$$

where S_0 is a Markov transition matrix, I is the identity matrix, and $r > 0$ is fixed. A Markov process X with rate matrix (5) can be realized by first constructing a Poisson process with rate r and jump times $\{T_k : k \geq 1\}$. The continuous-time process X is constant on the inter-jump time intervals $[T_k, T_{k+1})$, and

$$\mathbb{P}\{X_{T_{k+1}} = x' | X_{T_k} = x\} = S_0(x, x')$$

for $x, x' \in \mathbf{X}$ and $k \geq 0$, with $T_0 \equiv 0$. The assumption that $X_t = X_{T_k}$ for $t \in [T_k, T_{k+1})$ reflects the fact that we are only considering the load at the sampling times $\{T_k\}$.

The construction of S_0 is, of course, entirely dependent on the characteristics of the particular load. For TCLs, this construction is detailed in Section III-C. We assume throughout that it can be factored, respecting the nature/nurture constraints

$$S_0(x, x') = R_0(x, x'_u) Q_0(x, x'_n), \quad x, x' \in \mathbf{X} = \mathbf{X}^u \times \mathbf{X}^n \quad (6)$$

with $\sum_{x'_u} R_0(x, x'_u) = \sum_{x'_n} Q_0(x, x'_n) = 1$. Consequently,

$$\begin{aligned} \mathbb{P}\{X_{T_{k+1}}^n = x'_n | X_{T_k} = x\} &= Q_0(x, x'_n) \\ \mathbb{P}\{X_{T_{k+1}}^u = x'_u | X_{T_k} = x\} &= R_0(x, x'_u). \end{aligned}$$

An example of the transition matrix R_0 is postponed to (16).

Local control is based on the perturbation of nominal behavior, defined by a family of rate matrices $\{A_\zeta : \zeta \in \mathbb{R}\}$. The following structure is imposed:

$$A_\zeta = r[-I + S_\zeta] \quad (7)$$

in which S_ζ is a smooth (C^1) function of ζ and also factors

$$S_\zeta(x, x') = R_\zeta(x, x'_u) Q_0(x, x'_n), \quad x, x' \in \mathbf{X} = \mathbf{X}^u \times \mathbf{X}^n. \quad (8)$$

Observe that Q_0 remains fixed: we cannot modify the dynamics of nature.

Denote by $\mathcal{U}$ the power consumption as a function of state for an individual load: $\mathcal{U}(0, x_n) = 0$, and $\mathcal{U}(1, x_n) = \varrho$ (a positive value, independent of $x_n \in \mathbf{X}^n$).

2) *Mean-Field Models*: We describe here the construction of a mean-field model for a *homogeneous* population of loads. Analogous results for a heterogeneous population are summarized in Proposition 2.

The goal is to obtain a model of aggregate input-output behavior. In this article, the input is ζ , and the output is the aggregate power consumption

$$y_t^N := \frac{1}{N} \sum_{i=1}^N \mathcal{U}(X_t^i). \quad (9)$$

An approximate model is obtained from a large population limit, as $N \rightarrow \infty$. Analysis is complicated by the fact that the input is a function of aggregate measurements.

Suppose that a continuous reference signal r is given, and the goal is to achieve $y_t^N \approx r_t$ (perhaps after a transient period). These assumptions are imposed only for identification of the mean-field limit—there is no r in Fig. 1.

A fixed dynamic compensator is assumed given, defined by a pair of transfer functions with impulse responses denoted g and h , so that the input is specified by

$$\zeta_t^N = \int_0^t [g_{t-\tau} y_\tau^N + h_{t-\tau} r_\tau] d\tau, \quad t \geq 0. \quad (10)$$

The fact that the reference and impulse responses are fixed greatly simplifies analysis. For the control architecture considered in this article, we have $r_t \equiv U_t$, $h = h$ (the impulse response corresponding to M), and $g \equiv 0$.

Given a homogeneous collection of N loads, the empirical distribution at time t is denoted

$$\mu_t^N(x) := \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{X_t^i = x\}, \quad x \in \mathbf{X} \quad (11)$$

where $\mathbb{I}$ is the indicator function. Algebraic manipulations imply the representation

$$y_t^N = \sum_{x \in \mathbf{X}} \mu_t^N(x) \mathcal{U}(x). \quad (12)$$

That is, the output y_t^N is a linear function of the empirical distribution μ_t^N for each t . Most valuable is the approximate dynamics: Proposition 1 below presents conditions under which its evolution is approximated by the mean-field equation

$$\frac{d}{dt} \mu_t(x') = \sum_{x \in \mathbf{X}} \mu_t(x) \mathcal{A}_{\zeta_t}(x, x'), \quad x' \in \mathbf{X}.$$

In the theory of Markov processes and mean-field control, it is customary to interpret the right-hand side as vector-matrix multiplication, with μ_t interpreted as a row vector. This motivates the compact notation

$$\frac{d}{dt} \mu_t = \mu_t \mathcal{A}_{\zeta_t}. \quad (13)$$

The output (12) can then be interpreted as a dot-product. The squared error at time t is denoted $\|\mu_t^N - \mu_t\|^2 = \sum_{x \in \mathbf{X}} |\mu_t^N(x) - \mu_t(x)|^2$.

Proposition 1: Suppose that the following holds: the rate matrices $\{A_\zeta : \zeta \in \mathbb{R}\}$ are Lipschitz continuous in ζ , for each $N \geq 1$, the input is defined by (10), and that $\mu_0^N \rightarrow \mu_0$ in L_2 .

Then, the empirical distributions converge uniformly in L_2 , on any finite interval

$$\lim_{N \rightarrow \infty} \sup_{0 \leq t \leq T} \sum_x \mathbb{E}[\|\mu_t^N - \mu_t\|^2] = 0, \quad T > 0$$

where μ_t is defined in (13) with initial condition μ_0 , and input

$$\zeta_t = \int_0^t [g_{t-\tau} y_\tau + h_{t-\tau} r_\tau] d\tau, \quad y_t := \sum_x \mu_t(x) \mathcal{U}(x). \quad \square$$

The proof is found in the Appendix, in which the main step is the following Martingale representation [21, Ch. 4]:

$$\begin{aligned} \mu_t^N(x) &= \mu_0^N(x) + \sum_{x^-} \int_0^t \mu_{x^-}^N(x^-) \mathcal{A}_{\zeta_{x^-}^N}(x^-, x) \\ &\quad + N^{-1/2} M_t^N(x) \end{aligned} \quad (14)$$

where M_t is a Martingale whose variance is uniformly bounded in N on any finite interval $[0, T]$.

The quantity $y_t = \sum_x \mu_t(x) \mathcal{U}(x)$ is the average power for the mean-field limit, and the steady-state average power consumption for the nominal model is denoted $\bar{y}^0 = \sum_x \pi_0(x) \mathcal{U}(x)$ [recall definition of π_0 below (4)].

The smoothness assumption on S_ζ justifies the small-signal linear state-space model approximation

$$\frac{d}{dt}\Phi_t = A\Phi_t + B\zeta_t, \quad \gamma_t = C\Phi_t \quad (15)$$

in which $\Phi_t(k)$ approximates $\mu_t(x^k) - \pi_0(x^k)$ for $1 \leq k \leq d$, and the output γ_t is an approximation of $\tilde{y}_t = y_t - \bar{y}^0$.

The linear system parameters are easily identified: $A = \mathcal{A}_0^r$, and B, C^r are the column vectors of dimension $d = |\mathbf{X}|$

$$B_k = \sum_x \pi_0(x) \mathcal{A}_0'(x, x^k), \quad C_k = \mathcal{U}(x^k), \quad 1 \leq k \leq d$$

where $\mathcal{A}_0'$ is the derivative of $\mathcal{A}_\zeta$ at $\zeta = 0$, which by (7) becomes $\mathcal{A}_0' = rS_0'$.

C. Markov Models of TCLs

The standard model of a TCL is a first-order ODE with hysteresis control: the temperature at time t is denoted θ_t , taking values in the hysteresis interval $[\theta_{\min}, \theta_{\max}]$ (a full discussion of the ODE can be found in Appendix C-B). A finite state-space Markovian model can be obtained by binning temperature, and a smooth perturbation of the hysteresis control forms the “nurture dynamics.” An approach to the design of the nominal model is summarized in the following.

1) *Binning*: Recall $\mathbf{X}^u = \{0, 1\}$. A finite state space $\mathbf{X}^n$ is obtained through quantization of the interval $[\theta_{\min}, \theta_{\max}]$. For a given integer d , the interval $[\theta_{\min}, \theta_{\max}]$ is discretized into $d/2$ values as follows: $\mathbf{X}^n = \{\theta_{\min} + k\theta_\Delta : 0 \leq k \leq d/2 - 1\}$, where $\theta_\Delta = (\theta_{\max} - \theta_{\min})/(d/2 - 1)$ represents the temperature increments (step size) in the interval $[\theta_{\min}, \theta_{\max}]$.

We next describe the construction of a Markovian model. For consistency with (5), it is assumed that jumps occur only at the sampling times $\{T_k : k \geq 0\}$, with $T_0 = 0$ and $\{T_k : k \geq 1\}$ generated as the jump times of a Poisson process with rate r . This will be a reasonable approximation if $\theta_\Delta > 0$ is small.

2) *Nurture*: The nurture component X^u evolves on $\{0, 1\}$ with jump times restricted to $\{T_k : k \geq 1\}$ by design. The matrix R_0 takes the form

$$R_0(x, 1) = \begin{cases} 1 - p^\ominus(x_n), & x = (1, x_n) \\ p^\oplus(x_n), & x = (0, x_n) \end{cases}$$

$$R_0(x, 0) = \begin{cases} p^\ominus(x_n), & x = (1, x_n) \\ 1 - p^\oplus(x_n), & x = (0, x_n). \end{cases} \quad (16)$$

The switching probabilities $\{p^\oplus(x_n), p^\ominus(x_n)\}$ are designed to approximate standard hysteresis, meaning that each is approximately zero when x_n is far from the boundary of the interval $[\theta_{\min}, \theta_{\max}]$: additional details can be found in Appendix C-A.

3) *Nature*: The matrix Q_0 that models the dynamics of $\mathbf{X}^n$ is considered next. For each $x = (x_u, x_n) \in \mathbf{X}$, $x'_n \in \mathbf{X}^n$, recall that

$$Q_0(x, x'_n) = \mathbf{P}\{X_{T_{k+1}}^n = x'_n | X_{T_k} = x\}.$$

These probabilities can be estimated based on data gathered from physical experiments or simulation.

One approach is to compute the empirical mean based on $K \gg 1$ observations

$$\pi_Q(x, x'_n) := \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{I}(X_{T_k} = x, X_{T_{k+1}}^n = x'_n)$$

and obtain an empirical model, motivated by the Bayes rule

$$\widehat{Q}_0(x, x'_n) = \frac{\pi_Q(x, x'_n)}{\sum_{x''_n} \pi_Q(x, x''_n)}. \quad (17)$$

IV. HOMOGENEITY BY DESIGN

The transfer function G_ℓ in Fig. 2 was obtained from a version of a linear state-space model of the form (15). The resonance is found to be typical in previous studies, which is our motivation for introducing filtering techniques to suppress the resonance and restrict bandwidth of service for each load.

Local control design is performed in two stages.

- 1) The ideal case is considered, in which there is a large population of loads with identical characteristics. This is the setting of Proposition 1.
- 2) For a heterogeneous population, we take a given load and use the filter obtained in (i) based on the (false) assumption that the population is homogeneous. Filter design can be performed at the load, since it is a simple task to obtain the linear model locally based on input–output observations.

Justification of this approach requires an extension of Proposition 1 that incorporates heterogeneity: this is postponed to the end of the section, in Proposition 2.

A. Heterogeneous Population

The generalization of Proposition 1 for a heterogeneous population requires a refined model. We assume we have a doubly parameterized family of rate matrices $\{\mathcal{A}_\zeta^\vartheta : \zeta \in \mathbb{R}, \vartheta \in \mathbb{R}^m\}$, where the parameter ϑ represents the particular type of load.

It is assumed that the i th load has parameter ϑ^i , and that the sequence $\{\vartheta^i : i \geq 1\}$ is independent and identically distributed (i.i.d.) with marginal density denoted f_ϑ . Filtering is applied, with filter M^ϑ possibly different for each ϑ . To take this into account in a mean-field analysis, we require further notation.

An extension of (10) is defined by a family of impulse responses $\{g^\vartheta, h^\vartheta : \vartheta \in \mathbb{R}^d\}$. The input to a load with parameter ϑ is defined by

$$\zeta_t^{\vartheta, N} = \int_0^t [g_{t-\tau}^\vartheta y_\tau^N + h_{t-\tau}^\vartheta r_\tau] d\tau \quad (18a)$$

where again $y_t^N = N^{-1} \sum_{i=1}^N \mathcal{U}(X_t^i)$.

The definition of the empirical distributions requires the following extension: for each $t \geq 0$, $x \in \mathbf{X}$, and Borel set $S \subset \mathbb{R}^m$

$$\Lambda_t^N(x, S) := \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{X_t^i = x, \vartheta^i \in S\} \quad (19a)$$

$$\mu_t^N(x) := \Lambda_t^N(x, \mathbb{R}^m). \quad (19b)$$

Proposition 2 extends Proposition 1, establishing conditions under which (19a) is convergent, as $N \rightarrow \infty$. The limit Λ_t can be expressed in terms of a density λ_t , which evolves according to familiar dynamics

$$\begin{aligned} \frac{d}{dt}\lambda_t(x, \vartheta) &= \sum_{x^-} \lambda_t(x^-, \vartheta) \mathcal{A}_{\zeta_t}^\vartheta(x^-, x) \\ \Lambda_t(x, S) &= \int_{\vartheta \in S} \lambda_t(x, \vartheta) f_\vartheta(\vartheta) d\vartheta. \end{aligned} \quad (19c)$$

Proposition 2 requires the following assumptions:

Assumption 1:

- 1) The rate matrices $\{\mathcal{A}_\zeta^\vartheta : \zeta \in \mathbb{R}, \vartheta \in \mathbb{R}^m\}$ are continuously differentiable in $(\zeta, \vartheta) \in \mathbb{R}^{m+1}$ and Lipschitz continuous in ζ with Lipschitz constant independent of ϑ .
- 2) For each $T > 0$, the family of impulse responses $\{h_\tau^\vartheta : \vartheta \in \mathbb{R}^m, \tau \geq 0\}$ is uniformly bounded in both ϑ and $\tau \in [0, T]$, and continuously differentiable in ϑ .
- 3) The initial conditions $\{X_0^i : i \geq 1\}$ are independent of the parameters $\{\vartheta^i : i \geq 1\}$ and $\mu_0^N \rightarrow \mu_0$ in L_2 .

The proof of Proposition 2 is postponed to the Appendix.

Proposition 2: Suppose that Assumption 1 holds. Then, for any finite interval $[0, T]$, any $x \in \mathbb{X}$, and any open set $S \subset \mathbb{R}^m$

$$\begin{aligned} \lim_{N \rightarrow \infty} \sup_{0 \leq t \leq T} \mathbb{E} \left[|\Lambda_t^N(x, S) - \Lambda_t(x, S)|^2 \right] &= 0 \\ \lim_{N \rightarrow \infty} \sup_{0 \leq t \leq T} \sum_x \mathbb{E} \left[|\mu_t^N(x) - \mu_t(x)|^2 \right] &= 0 \end{aligned}$$

where $\{\Lambda_t^N, \Lambda_t, \mu_t^N, \mu_t\}$ are defined in (19), in which the differential equation (19c) defining λ_t is initialized with $\lambda_0 = \mu_0 \times f_\vartheta$, and

$$\zeta_t^\vartheta = \int_0^t [g_{t-\tau}^\vartheta y_\tau + h_{t-\tau}^\vartheta r_\tau] d\tau \quad (20a)$$

$$y_t = \sum_x \mu_t(x) \mathcal{U}(x), \quad \mu_t(x) = \int \lambda_t(x, \vartheta) f_\vartheta(\vartheta) d\vartheta. \quad (20b)$$

□

Let G_ℓ^ϑ denote the transfer function for the linearized dynamics for the model with parameter ϑ . The proposition easily justifies the “optimistic” control strategy, in which each inverse filter M^ϑ is designed based on G_ℓ^ϑ under “*the (false) assumption that the population is homogeneous*” (italics refer to quoted text from the start of Section IV).

Corollary 1: Suppose that $\varepsilon > 0$ and a frequency interval $\Omega_{\text{des}} \subset \mathbb{R}_+$ is given, and the local filters satisfy

$$|M^\vartheta(jf)G_\ell^\vartheta(jf) - 1| \leq \varepsilon, \quad \text{for each } f \in \Omega_{\text{des}}, \vartheta \in \mathbb{R}^m.$$

Consider the input–output model with input ζ and output y defined by the pair of equations (19c) and (20b); let G_ℓ denote its linearization at $\zeta \equiv 0$.

If, in addition, the assumptions of Proposition 2 hold, we then have

$$|G_\ell(jf) - 1| \leq \varepsilon, \quad \text{for each } f \in \Omega_{\text{des}}.$$

Proof: Let $y_t(\vartheta) = \sum_x \lambda_t(x, \vartheta) \mathcal{U}(x)$, so that

$$y_t = \int y_t(\vartheta) f_\vartheta(\vartheta) d\vartheta.$$

The small-signal linearization of the input–output system with input ζ_t and output $y_t(\vartheta)$ has transfer function $M^\vartheta G_\ell^\vartheta$ by construction. We thus obtain

$$G_\ell(jf) - 1 = \int \{M^\vartheta(jf)G_\ell^\vartheta(jf) - 1\} f_\vartheta(\vartheta) d\vartheta.$$

Jensen’s inequality completes the proof. □

Section V contains summaries of results from experiments to illustrate Proposition 2 and its corollary. The experiments surveyed in Section VI were designed in part to investigate the impact of unmodeled dynamics in the control architecture Fig. 1.

V. STOCHASTIC SIMULATIONS

The inverse filter design focused on a linearized model. Here, we consider a heterogeneous collection of TCLs and see whether the input–output behavior approximates what is predicted by a linear approximation.

A. Markovian Load Models and Filter Design

We begin with the three ingredients in the specification of a simulation model for an individual load.

1) *Nurture:* The matrix R_0 was chosen of the form (16) in which the switching probabilities $\{p^\oplus(x_n), p^\ominus(x_n)\}$ differ depending on the load. Details are found in the Appendix.

The local randomized control was chosen to be the myopic design of [8]: for $x'_u \in \{0, 1\}$ and $x \in \mathbb{X}$

$$R_\zeta(x, x'_u) = \frac{1}{\kappa_\zeta(x)} R_0(x, x'_u) \exp(\zeta \mathcal{U}(x'_u)) \quad (21)$$

in which $\kappa_\zeta(x)$ is the normalizing constant defined so that $\sum_{x'_u} R_\zeta(x, x'_u) = 1$ for each x .

This policy influences the load to increase power consumption at time t when $\zeta_t > 0$, and less power when $\zeta_t < 0$.

2) *Nature:* For each load, the system was simulated using the policy defined by R_0 (an approximation of hysteresis). Data collected from a long simulation were collected, and the transition matrix Q_0 was estimated via Monte Carlo as described in Section III-C.

3) *Filter Design:* Recall from the introduction and Fig. 1 that it is assumed that the BA broadcasts a common signal U_t to each load participating in demand dispatch, and recall that this is a simplification of current practice.

Referring once more to Fig. 1, it is reasonable to assume that the services provided from flexibility of pool pumps are very different from those provided from a residential water heater or refrigerator. It is argued in [6] that a formalism to justify this assumption can be found by classifying on/off loads according to their nominal cycling patterns. A crude summary is obtained on consideration of tracking a square wave: a population of pool pumps can track a symmetric square wave of period 24 h perfectly by turning off and on in unison. The same is approximately true for a population of residential refrigerators, but with a period of approximately 1 h.

This motivates the introduction of bandpass filters, designed based on the spectral characteristics of the load. It may be that the BA performs filtering, as done today at PJM.

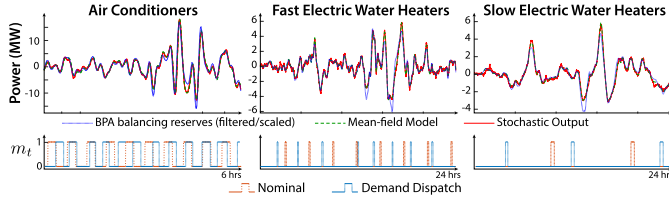


Fig. 5. Open-loop tracking with 40 000 heterogeneous TCLs in each class. The bottom plots show the ON/OFF-state m for a typical load of each of the three classes.

For consistency of notation, it is assumed in the remainder of this article that this additional filtering is performed at the load, so that U_t remains a common command signal, broadcast to every asset. Butterworth filters are adopted in the numerical results that follow.

An adjustment of notation is required: Constructed at load with parameter ϑ is a Butterworth filter $M^{\text{BW}\vartheta}$ and approximate inverse filter $M^{\text{inv}\vartheta}$ designed based on the methodology surveyed in Section IV. The transfer function M^ϑ appearing in Fig. 3 is the product $M^\vartheta = M^{\text{BW}\vartheta} M^{\text{inv}\vartheta}$.

B. Open-Loop Tracking

1) *Summary of Load Classes and Subclasses:* We present here details regarding the diversity of loads involved. The precise details of the simulation environment are summarized in Appendix C; Tables I and II may be found there.

Three classes of TCLs were considered: residential air conditioners (ac), small electric water heaters with faster cycle times (f-WH), and large electric water heaters with slower cycle times (s-WH). Twenty different subgroups were obtained for each TCL class, through uniform sampling of the values in Table I. Each subgroup contains 2000 loads, resulting in a total of 40 000 loads in each TCL class.

The choice of bandpass filters was based on the nominal period of each of the load classes. The parameters of the Butterworth filters are summarized in Table II.

2) *Tracking BPA's Balancing Reserves:* The balancing reserves deployed (BRD) from the Bonneville Power Administration (BPA) were used as a reference signal. A single typical windy day, 19 February 2016, was chosen.

Fig. 5 shows *open-loop* tracking for each of three TCL classes (for the case of ac, the plot shows only 6 h during the day). Capacity estimates were obtained using the approximations in [10] and [23]. The ac class tracks a signal that is approximately 20% of the power capacity of the aggregate (see $\bar{q}_{\text{tot}}$ in Table II for power capacity). We estimate that the ac aggregation can track a signal of ± 100 MW while maintaining reasonable tracking error. The other two plots, featuring the water heating loads, represent about 60% of the power capacity of the collection of water heaters. We observe clipping of the reference signal when it reaches power capacity limits.

It was found that cycling of TCLs was increased by about 5% from nominal.

The tracking accuracy is remarkable for a one-way communication architecture from the BA to the loads. These results might be anticipated from Proposition 2 and its corollary, but

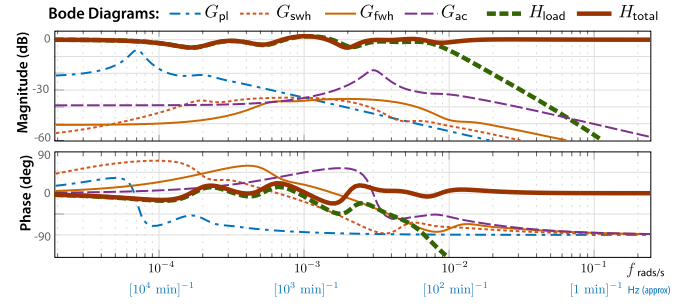


Fig. 6. Macro load dynamics. H_{load} is the transfer function of the sum of linearized dynamics from a heterogeneous collection of pools, water heaters, and ACs. It has an approximately flat gain and phase response in the bandwidth of interest (for services from flexible loads).

remember that the filters $\{M^\vartheta\}$ are based on a linear model, while the actual aggregate dynamics are far from this linear ideal.

VI. GRID-LEVEL SIMULATIONS

The results from grid-level simulations are surveyed in this section, based on the architecture shown in Fig. 1. Performance of tracking and disturbance rejection was investigated, along with a cost analysis. The impact of unmodeled dynamics was also investigated through simulations.

A. Simulation Model

Fig. 5 not only shows good open-loop tracking but also demonstrates that the mean-field model is a highly accurate approximation of the stochastic system. For this reason, *throughout this section simulations are conducted using the mean-field model.*

The simulations were based on the grid model shown in Fig. 1. Details are summarized as follows.

- 1) The grid-level transfer functions G_p and G_c seen in Fig. 1 are given in (2) and (3).
- 2) The disturbance D entering the grid (modeled as an additive input disturbance as shown in Fig. 1) was taken to be the BPA BRD data from 19 to 23 February 2016. This modeling choice reflects the fact that the grid operator at BPA wishes the “actuators” in its domain to track accurately the BRD.
- 3) Only four load classes were considered in the *actuator block*: three TCLs and a collection of 1-kW pool pumps with 12-h cleaning each day. To reduce complexity, a single TCL model was chosen as representative of the 20 among the subgroups considered in Section V, so that in total only four values of ϑ were required in simulations. The mean-field dynamics within a group are thus

$$\frac{d}{dt}\mu_t^\vartheta = \mu_t^\vartheta \mathcal{A}_{\zeta_t}^\vartheta, \quad \zeta_t^\vartheta = [h^\vartheta * U]_t, \quad t \geq 0. \quad (22)$$

- 4) The mean-field model output $y_t^\vartheta := \sum_x \mu_t^\vartheta(x) \mathcal{U}(x)$ must be scaled to reflect the number of loads in the group. The simulations that follow scale this quantity, based on the following assumptions: 1 million ACs, 5 million f-wh, 5 million s-wh, and a large number of pools (this number was taken as a parameter in this study).

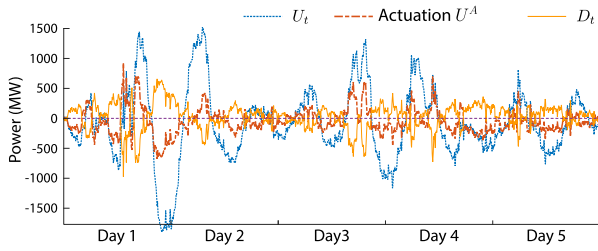


Fig. 7. Closed-loop tracking with residential air conditioners, electric water heaters, pool pumps, and ideal actuators. Actuation from loads is interpreted as virtual energy storage. U is the output of the compensator G_c , while U^A is the aggregate response of all the actuators (ACs + s-WH + f-WH + Pools + ideal resources).

The nonlinear control system was simulated in continuous time using Simulink, requiring parallel simulation of (22) for each value of ϑ . This was implemented using four linear parameter-varying (LPV) Simulink blocks run in parallel, with ζ_t^ϑ the parameter and μ_t^ϑ the d -dimensional state for each ϑ .

The linearized mean-field model of each load class is denoted G_{ac} , G_{fwh} , G_{swh} , and G_{pl} , and the respective filters (inverse $\times$ bandpass) by M_{ac} , M_{fwh} , M_{swh} , and M_{pl} . The linear model of the aggregate dynamics of all the loads is defined by the sum

$$H_{load} = M_{ac}G_{ac} + M_{fwh}G_{fwh} + M_{swh}G_{swh} + M_{pl}G_{pl}. \quad (23)$$

The Bode plot for H_{load} is shown in Fig. 6. The rapid decline in the magnitude plot beyond $f = 10^{-2}$ rads/s is due to the inherent bandwidth constraints of the loads. Hence, the actuation is augmented with an ideal resource $G_a \equiv 1$. A high-pass filter M^{HP} was designed with bandwidth beyond $f = 10^{-2}$ rads/s, so that the introduction of this resource flattens the Bode plot. The total response is modeled by the transfer function $H_{total} = H_{load} + M^{HP}G_a$, whose Bode plot is also shown in Fig. 6.

The actuation obtained from G_a might come from batteries, responsive generators, or fast responding loads that provide accurate tracking. The time-scales of ancillary service from these resources are predominately in the range of primary control (e.g. *droop*).

Is this a perfect virtual battery? Of course there are imperfections: the nonlinear dynamics have been linearized for the sake of control design, but the system dynamics remain nonlinear. Moreover, the Bode plot for the linearized dynamics with transfer function H_{total} is not entirely flat in magnitude or phase.

These shortcomings are no different than what would be expected for a generator providing balancing service [25], or a realistic (and imperfect) battery system. Moreover, we will see in the remainder of this section that global feedback compensates for imperfections in actuation.

B. Closed-Loop Performance

With this mix of TCL loads, 4 million pools will be required to accurately track the BRD signal as the disturbance entering the grid! In the results surveyed here, we cap the number of pools at 1 million (the approximate number of residential pools in Florida, with even more in California). The maximum load is thus 1 GW, and the average load is 500 MW,

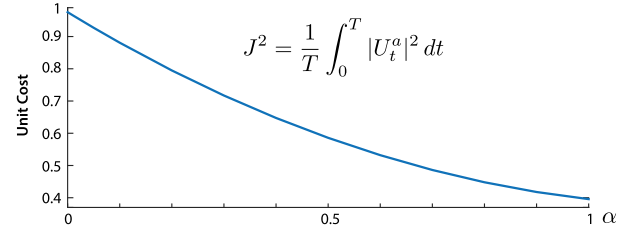
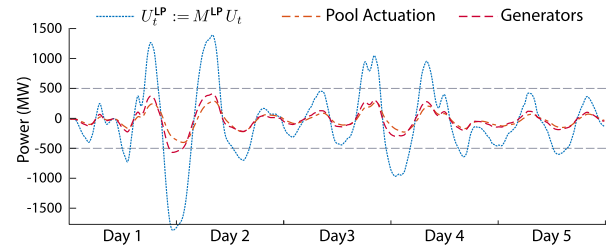


Fig. 8. Normalized cost as a function of capacity from ac loads.

so that the pools can at best track signals of ± 500 MW. Tracking was poor when the BRD signal exceeded this range.

Other resources such as commercial water chillers could be added to increase capacity at low frequencies. Instead, in the next set of simulations, the pools were augmented with a single 1-GW generator. This was modeled through the introduction of an additional ideal actuator

$$H_{total} = H_{load} + M^{HP}G_a + \frac{1}{4}M^{LP}G_a$$

in which the second-order low-pass filter M^{LP} has unity gain at low frequencies, chosen identical to the Butterworth filter adopted for the pool loads. The scaling of 1/4 is introduced so that the response from the ideal low-frequency actuators is commensurate with the pools. The resulting Bode plot is no longer flat—its gain below 10^{-4} rads/s is approximately half of the gain above 10^{-3} rads/s.

As a result of the gain variations in the linearization and the nonlinearities inherent in mean field dynamics, the open-loop tracking performance will suffer, especially when the BRD signal takes on large values. While imperfect, the performance is still better than what is received from many generation units (for example, see [25, Fig. 10]).

The plots on the left-hand side of Fig. 7 show closed-loop behavior over 5 days. The aggregate response from all the actuators, U^A , is approximately the negative of the BRD, so that the frequency deviation is tightly controlled: it remains within the range 59.993–60.007 Hz over the five-day period.

The plot on the right-hand side of Fig. 7 shows the filtered control signal $U_t^{LP} := M^{LP}U_t$ along with two responses: from the collection of pools, and from the 1-GW generator. The response of the pools nearly matches the response from the ideal generator.

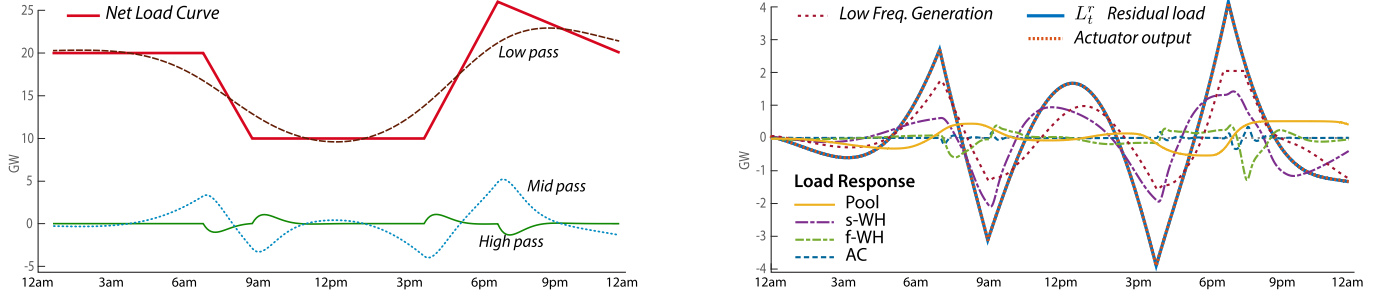


Fig. 9. Left: Hypothetical CAISO net load over one day in 2020, and its frequency decomposition. Right: “Residual Load” = “Net Load” – “Low Pass” is tracked nearly perfectly. The introduction of demand dispatch alongside generation reduces needed generation capacity by at least 5 GW.

C. Time-Varying Capacity

The time-varying nature of many commercial and residential loads is an issue of concern. For example, the number of air conditioners that are in operation, and hence available for ancillary service, is low during the early morning hours and peaks during the late afternoon [39].

The experiments were conducted in which the gain of the response of the ACs was amplified/attenuated using a time-varying gain function

$$g(t) = 1 - 0.5 \sin(f_d t), \quad t \geq 0$$

in which $f_d = 727 \times 10^{-7}$ rads/s corresponds to a 24-h period. All the other resources were left the same as the simulation setting of Section VI-B. The reader is referred to [31] for plots of the response of the aggregate of loads in similar experiments. The time variations result in significant change in response from ACs when compared with nominal, yet the grid frequency remains within [59.993, 60.007] Hz due to grid-level feedback (referring again to Fig. 1).

D. Resource Availability and Cost

Following installation of equipment to enable demand dispatch, the operating cost is essentially zero. Consumers may require incentives to participate (e.g., FPL and Austin Energy provide credits to participating residents), but they will also receive some guarantees regarding constraints on QoS and potential costs from additional cycling of equipment.

The benefit of demand dispatch from low-frequency services such as residential pools is clear: 1 million pools might serve as a substitute for a 500-MW generator. Following the initial investment (usually in \$B), a generator requires fuel, maintenance, and staff. The loads provide accurate regulation service without any of these operating costs.

What about higher frequency ancillary services? To investigate the value of the highest frequency services from demand dispatch, we consider a parameterized family of models in which the contribution from air conditioners is varied according to the fraction $\alpha \in [0, 1]$. The remaining $1 - \alpha$ of regulation is obtained from ideal actuation from batteries or other sources. Denote the output of the ideal actuators by $\{U_t^a\}$. The total ideal actuation is defined by the sum

$$U_t^a = [M^{\text{HP}} G_a + (1 - \alpha) M_{\text{ac}}^{\text{BP}} G_a] U_t.$$

Recall $G_a \equiv 1$ in these experiments. The second component is thus $(1 - \alpha) M_{\text{ac}}^{\text{BP}} U_t$, which is intended to replace the lost service from the ACs.

The mean-square cost of the closed-loop system is defined as

$$J^2 = \frac{1}{T} \int_0^T |U_t^a|^2 dt. \quad (24)$$

This is similar to the “mileage” metric used for ancillary service resources such as batteries. Fig. 8 shows a plot of this cost as a function of α for T corresponding to one day; $\alpha = 1$ corresponds to the simulation setting of Section VI-B. The total cost is reduced by more than 50% when $\alpha = 1$ when compared with $\alpha = 0$.

The cost is much higher for intermediate values of α when the inverse filter is not used to construct ξ^{ac} .

E. Ramp Services

The plot on the right in Fig. 9 shows a stylized “duck curve” representing the net load (load – renewables) at CAISO; the shape is a result of significant penetration of solar energy in California. The plot reflects approximately 10 GW of solar power at peak.

The 15-GW ramp observed between 3 P.M. and 6 P.M. is of concern today. The ramp can be smoothed by first scheduling generation to track a low-frequency component of the net load—denoted “low pass” in the figure. The remaining two zero-energy signals shown can be tracked using a combination of resources—batteries, responsive generators, and demand dispatch.

The mid-pass signal remains substantial—a range of ± 5 GW. This signal could be provided using gas turbine generators, but a total capacity of 10 GW would be required. This value can be reduced significantly by applying the same techniques used to address the balancing reserves signal.

Let L_t^r denote the residual load, defined as “Net Load” – “Low Pass.” This is plotted on the right in Fig. 9, where it is seen that it takes on values approaching ± 4 GW. The capacity from loads in the previous set of experiments was insufficient to track this signal. The capacity from TCLs was doubled, so that the simulation was based on 10 million s-WH, 10 million f-WH, and 2 million ACs. It included 1.2 million pools (the approximate number of pools in California), and also ± 2 GW of low-frequency regulation that might come

from generation or demand dispatch from other loads such as water chillers and water pumping (a significant load in California).

The plots of the power deviation from TCLs shown on the right in Fig. 9 are significant, even though temperature constraints at the load are not violated. The variation in power consumption of s-WH and pools helps address the “mid pass” signal shown on the left of Fig. 9, whereas the “high pass” component is serviced by the f-WH and ac power consumption. The residual load and aggregate actuation match nearly perfectly.

VII. CONCLUSION

It is exciting to see how, through distributed control design, heterogeneous flexible loads can coordinate to smooth out enormous shocks to the grid. From the grid-operator’s perspective, a collection of heterogeneous loads can be as valuable as a multi-GW battery.

The theory and simulations presented in this article show the success of the demand dispatch architecture with minimal measurements required from the loads; the grid operator will require estimates of capacity from flexible load aggregations, but does not require continuous power readings from individual loads.

The hardware requirements for smart loads (e.g. NEST thermostats or smart fridges) should be minimal, with communication based on a combination of internet and power line communication (the original approach of FP&L [1]).

Two issues require further attention. First is the role of the “perfect actuators” supplying regulation at time scales of tens of seconds and faster (the timescale of today’s primary control). Can loads assist with this service and bolster synthetic inertia? Analysis in [32] suggests that this could bring risk in terms of stability, but this may depend on other elements of the grid architecture (e.g., the number and size of synchronous generators) [24], [43].

The impact of the time-varying nature of many loads deserves further study. For example, the nominal load from commercial and residential air conditioning is roughly periodic over a typical week, and its magnitude changes slowly depending on the weather. The results summarized in Section VI-C offer significant hope in terms of system stability. It is conjectured that periodicity is a benefit in regions with significant solar energy, since demand is in harmony with supply.

The impact of distribution is also of interest: will voltage constraints in the distribution network limit capacity from virtual energy storage? Conversely, what is the potential for application of virtual energy storage for providing voltage support and grid support simultaneously?

APPENDIX A MEAN FIELD MODELS

Proposition 1 and Proposition 2 are based on the versions of the Martingale representation (14) and Grönwall’s inequality in this simple form.

Lemma 1: Let z be a nonnegative and continuous function on an interval $[0, T]$. Suppose that the following bound holds

for constants $\alpha, \beta > 0$:

$$z_t \leq \alpha + \beta \int_0^t z_r dr, \quad 0 \leq t \leq T.$$

Then $z_t \leq \alpha e^{\beta t}$ for $0 \leq t \leq T$.

Proof of Proposition 1: Let $\tilde{\mu}_t^N = \mu_t^N - \mu_t$ and $\|\tilde{\mu}_t^N\|_1 = \sum_x |\tilde{\mu}_t^N(x)|$. We apply Grönwall’s inequality using

$$z_t^N = \max_{r \leq t} \mathbb{E} \left[\|\tilde{\mu}_r^N\|_1 \right].$$

We begin with (14)

$$\begin{aligned} \tilde{\mu}_t^N(x) &= \tilde{\mu}_0^N(x) + \sum_{x^-} \int_0^t \tilde{\mu}_r^N(x^-) \mathcal{A}_{\zeta_r}(x^-, x) \\ &\quad + \sum_{x^-} \int_0^t \mu_r^N(x^-) \{ \mathcal{A}_{\zeta_r^N}(x^-, x) - \mathcal{A}_{\zeta_r}(x^-, x) \} dr \\ &\quad + N^{-1/2} M_t^N(x) \end{aligned}$$

where M_t is a Martingale whose variance is uniformly bounded in N and t , and $\zeta_t = \int_0^t g_{t-\tau} y_\tau d\tau$.

Because μ_r^N is a pmf, we can bound the second term

$$\begin{aligned} &\sum_{x^-} \int_0^t \mu_r^N(x^-) \{ \mathcal{A}_{\zeta_r^N}(x^-, x) - \mathcal{A}_{\zeta_r}(x^-, x) \} dr \\ &\leq \max_{x^-, x} \int_0^t |\mathcal{A}_{\zeta_r^N}(x^-, x) - \mathcal{A}_{\zeta_r}(x^-, x)| dr \\ &\leq L \int_0^t |\zeta_r^N - \zeta_r| dr \end{aligned}$$

where L is a Lipschitz constant for the family of rate matrices. We also have from the definitions

$$\zeta_r^N - \zeta_r = \sum_x \int_0^r g_{t-\tau} \tilde{\mu}_\tau^N(x) \mathcal{U}(x) d\tau$$

giving $\mathbb{E}[|\zeta_r^N - \zeta_r|] \leq \|g\|_{[0,T]} \|\mathcal{U}\|_\infty z_r^N$, where $\|g\|_{[0,T]} = \int_0^T |g(\tau)| d\tau$ and $\|\mathcal{U}\|_\infty = \max_{x_u} |\mathcal{U}(x_u)|$.

It follows that there is a finite constant β such that

$$\|\tilde{\mu}_t^N\|_{L_2} := \sqrt{\mathbb{E} \left[\|\tilde{\mu}_t^N\|_1^2 \right]} \leq \|\tilde{\mu}_0^N\|_{L_2} + \varepsilon_t^N + \beta \int_0^t z_s^N ds$$

where $\varepsilon_t^N = N^{-1/2} \|M_t^N\|_{L_2}$. Letting ε^N denote the maximum of ε_t^N over $[0, T]$ gives

$$z_t^N := \max_{r \leq t} \mathbb{E} \left[\|\tilde{\mu}_r^N\|_1 \right] \leq \|\tilde{\mu}_0^N\|_{L_2} + \varepsilon^N + \beta \int_0^t z_s^N ds.$$

It follows from Lemma A.1 that z_T^N vanishes as $N \rightarrow \infty$, which completes the proof. $\square$

Proof of Proposition 2: To prove the proposition, we restrict to open sets satisfying $S \subset \{\vartheta \in \mathbb{R}^m : \|\vartheta - \vartheta_0\| < \varepsilon_0\}$, with $\vartheta_0 \in \mathbb{R}^m$ arbitrary. It is enough to establish the bound

$$\limsup_{N \rightarrow \infty} \mathbb{E} \left[\left\{ \tilde{\Lambda}_t^N(x, S) \right\}^2 \right] = o(\varepsilon_0), \quad \tilde{\Lambda}_t^N = \Lambda_t^N - \Lambda_t. \quad (25)$$

The proof begins with another Martingale representation, this time for a single load. For each i, t , let δ_t^i denote the point mass supported at $X_t^i \in \mathbf{X}$. That is, $\delta_t^i(x) = \mathbb{I}\{X_t^i = x\}$

for each x . The following may be interpreted as a special case of (14) in which $N = 1$:

$$\delta_t^i(x) = \delta_0^i(x) + \sum_{x^-} \int_0^t \delta_r^i(x^-) \mathcal{A}_{\zeta_r^N}^{\vartheta^i}(x^-, x) dr + M_t^i(x)$$

where, for any $N \geq i$, the stochastic process $\{M_t^i : t \geq 0\}$ is a Martingale with respect to the filtration

$$\mathcal{F}_t^N = \sigma\{X_r^j : j \leq N, r \leq t, \vartheta^k : k \geq 1\}.$$

The special assumption on S combined with (19c) give

$$\Lambda_t(x, S) = \lambda_t(x, \vartheta_0) v_\theta(S) + o(\varepsilon_0)$$

where $v_\theta(S) = \int_{\vartheta \in S} f_\theta(\vartheta) d\vartheta$

and the ratio $o(\varepsilon_0)/\varepsilon_0$ vanishes with ε_0 , uniformly for $t \in [0, T]$. On integrating the ODE defining the density

$$\Lambda_t(x, S) = \Lambda_0(x, S) + \sum_{x^-} \int_0^t \Lambda_r(x^-, S) \mathcal{A}_{\zeta_r^0}^{\vartheta_0}(x^-, x) dr + o(\varepsilon_0). \quad (26)$$

For $N \geq 1$, we apply the Martingale representation of δ_t^i and the definition of Λ_t^N to obtain

$$\begin{aligned} \Lambda_t^N(x, S) &= \frac{1}{N} \sum_{i=1}^N \delta_t^i(x) \mathbb{I}\{\vartheta^i \in S\} \\ &= \Lambda_0^N(x, S) \\ &\quad + \frac{1}{N} \sum_{i=1}^N \left\{ \sum_{x^-} \int_0^t \delta_r^i(x^-) \mathcal{A}_{\zeta_r^N}^{\vartheta^i}(x^-, x) dr \right\} \mathbb{I}\{\vartheta^i \in S\} \\ &\quad + M_t(x, S) \end{aligned}$$

where the definition of $\{M_t(x, S)\}$ is self-evident. For fixed x and S , it is a Martingale with respect to the filtration $\mathcal{F}_t^N$.

As in the proof of Proposition 1, the next step is to replace ζ_r^N by $\zeta_r^{\vartheta_0}$ in the rate matrix. We do this and also replace ϑ^i with its approximation ϑ_0 to obtain with the help of (26)

$$\begin{aligned} \tilde{\Lambda}_t^N(x, S) &= \tilde{\Lambda}_0^N(x, S) \\ &\quad + \sum_{x^-} \int_0^t \tilde{\Lambda}_r^N(x^-, S) \mathcal{A}_{\zeta_r^{\vartheta_0}}^{\vartheta_0}(x^-, x) dr \\ &\quad + \sum_{x^-} \int_0^t \Lambda_r^N(x^-, S) \left\{ \mathcal{A}_{\zeta_r^N}^{\vartheta^i}(x^-, x) - \mathcal{A}_{\zeta_r^{\vartheta_0}}^{\vartheta_0}(x^-, x) \right\} dr \\ &\quad + M_t(x, S) + o(\varepsilon_0) \end{aligned}$$

where the final error term depends on t and N , but is uniformly bounded by a fixed constant time ε_0 .

The remainder of the proof of (25) follows from Grönwall's inequality as in Proposition 1, with the new definition

$$z_t^N = \max_{r \leq t} \mathbb{E} \left[\left(\sum_x |\tilde{\Lambda}_r^N(x, S)| \right)^2 \right]^{1/2}.$$

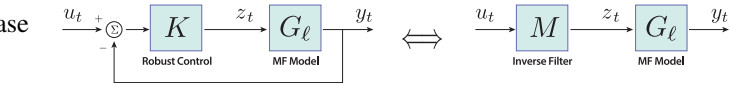


Fig. 10. Feedback control system for inverse filter design.

APPENDIX B FILTER DESIGN AND ROBUST CONTROL

Here, we consider a homogeneous population. Let G_ℓ denote the transfer function for the linearized mean-field model, corresponding to a state-space model of the form (15)

$$G_\ell(s) = C[sI - A]^{-1}B.$$

To simplify notation, we suppress dependency on ϑ , letting M denote the prefilter that appears in Fig. 3, with impulse response h .

The goal of filter design is based on a prespecified frequency interval $\Omega_{\text{des}} \subset \mathbb{R}_+$. The product $M(jf)G_\ell(jf)$ should be approximately unity for $f \in \Omega_{\text{des}}$ and vanish for frequencies far from this interval.

Application of robust control techniques is made possible through the change in variables

$$M = \frac{1}{1 + KG_\ell} K$$

where K is a transfer function to be designed. The *sensitivity function* and *complementary sensitivity function* are denoted

$$S := \frac{1}{1 + KG_\ell}, \quad T := 1 - S.$$

Motivation is provided in the feedback loop shown in Fig. 10, for which the transfer function from u_t to y_t is T . There are standard techniques to design K such that T is nearly unity in the frequency interval Ω_{des} and small for frequencies far from this interval.

Our criterion for choosing K includes the following three objectives:

$$\begin{aligned} S(jf) &\approx 0, & f \in \Omega_{\text{des}} \\ T(jf) &\approx 0, & f \notin \Omega_{\text{des}}, \text{ (equivalently, } T(jf) \approx 1) \end{aligned}$$

while maintaining reasonable bounds on $|M(jf)|$ for all f .

We use the mixed-sensitivity synthesis method to obtain a solution [28]. This requires three transfer functions (W_1, W_2, W_3) that serve as weights for the respective transfer functions (S, M, T). For any transfer function K , we define $(Q_1, Q_2, Q_3) = (W_1 S, W_2 M, W_3 T)$.

For any transfer function Z , denote $\|Z\|_{H_\infty} = \sup_f |Z(jf)|$. The objective function for optimization is

$$q(K) = \max_{1 \leq i \leq 3} \|Q_i\|_{H_\infty}.$$

The mixed-sensitivity synthesis method finds the transfer function that minimizes $q(K)$ over all proper transfer functions K . The optimizer can be obtained numerically, using the *mixsyn* command in MATLAB [5].

□

TABLE I

TCL PARAMETERS FOR ACS AND WATER HEATERS

Par.	AC	Fast WH	Slow WH
θ^{set}	18–22	48–52	48–52
δ	0.8–1	2.95–3	3.95–4
Θ^a	30–34	19–21	19–21
RC	3.5–4.5	30–36	67–73
ϱ	14/2.5	5/1	5/1
σ_W^2	10^{-6}	10^{-6}	10^{-6}

TABLE II

LOAD DYNAMICS AND POWER CHARACTERISTICS. BW REFERS TO BANDWIDTH OF BUTTERWORTH BANDPASS FILTERS; MAX AND AVERAGE POWER $\bar{Q}_{\text{tot}}$, $\bar{Q}_{\text{tot}}$ IN MW FOR 40 000 LOADS

Load	Period	BW (cyc/hr)	$\bar{Q}_{\text{tot}}$	$\bar{Q}_{\text{tot}}$
AC	20min–1hr.	[1, 1/0.2]	97	224
f-WH	2–4 hrs.	[1/3, 1/0.5]	11	200
s-WH	8–12 hrs.	[1/9, 1]	8.5	200
Pools	24 hrs.	[1/24, 1/3]	20	40

APPENDIX C

DETAILS OF NUMERICAL EXPERIMENTS

We first discuss how the matrix S_0 was defined for any of the on–off loads considered, and then explain how Q_0 was obtained for various TCL models. Here, we opted for simulation to fit a Markovian load model, since we did not have access to data for each class of TCLs.

Given a temperature set-point θ^{set} and a dead-band range δ , the TCL control is designed to maintain the temperature within these bounds: $\Theta_t \in [\theta^{\text{set}} - \delta/2, \theta^{\text{set}} + \delta/2]$ for each t . In each case considered, this interval is quantized into $d/2 = 20$ values. Consequently, $|X| = d = 40$.

The nominal transition matrix S_0 in (6) involves two maps: R_0 and Q_0 .

A. Nurture

To simplify the exposition, we restrict to heating loads. The details of constructing the nurture map R_0 are provided in Section III-C2. Here, we discuss the missing ingredients, i.e., the construction of the switching probabilities.

The construction begins with the specification of two cumulative density functions (CDFs), denoted $F^\oplus(x_n)$ and $F^\ominus(x_n)$. The switching probabilities are defined by

$$\begin{aligned} p^\oplus(x_n) &= F^\oplus(x_n) - F^\oplus(x_{n-1}) \\ p^\ominus(x_n) &= F^\ominus(x_n) - F^\ominus(x_{n-1}) \end{aligned}$$

where $x_n, x_{n-1} \in X^n$ for $0 \leq n \leq d/2 - 1$.

In the experiments surveyed here, the CDFs are specified by three parameters: $\theta_0^\ominus \in [\theta_{\min}, \theta_{\max}]$, $\nu \in (0, 1)$ and $\kappa > 1$. We then take

$$\begin{aligned} F^\ominus(x_n) &= (1 - \nu) \frac{[x_n - \theta_0^\ominus]_+^\kappa}{[\theta_{\max} - \theta_0^\ominus]^\kappa} \\ F^\oplus(x_n) &= 1 - F^\ominus(\theta_{\max} + \theta_{\min} - x_n), \quad x_n \in X^n \end{aligned}$$

with $[x]_+ := \max(0, x)$ for $x \in \mathbb{R}$; $\theta_0^\ominus = \theta_{\min}$, $\nu = 0.8$, and $\kappa = 4$.

B. Nature

Construction of Q_0 for a TCL began with the standard ODE model

$$\frac{d}{dt} \Theta_t = -\frac{1}{RC} (\Theta_t - \Theta_t^a + \theta^g m_t) + W_t \quad (27)$$

in which Θ_t is the internal temperature, $m_t \in \{0, 1\}$ represents the power mode, and W_t models disturbances (chosen Gaussian with zero mean and variance σ_W^2 found in Table I). The parameters are described as follows: Θ_t^a is the ambient temperature, C is the thermal capacitance, R is the thermal resistance, ϱ is the *coefficient of performance*, and $\theta^g = R\rho_{\text{tr}}$, where ρ_{tr} is the energy transfer rate; ρ_{tr} is positive for TCLs providing cooling, and negative otherwise.

The ODE (27) was simulated and sampled according to a Poisson process of rate r . At each sampling instant T_k , the randomized policy R_0 was used to determine the next power mode. Hence, m_t remains constant on each inter-sampling interval $[T_k, T_{k+1})$, and we let $X_k^u = m_{T_k-} = m_{T_{k-1}}$ for each k .

The data were collected to obtain the estimate $\hat{Q}_0$ defined in (17), which were then used to define Q_0 in the experiments reported in this article.

C. Details of Parameter Values

Table I displays the range of TCL parameters for air conditioners and electric water heaters (a subset of those surveyed in [33]). The temperature parameters are in degree Celsius. The value of RC is in units of time (hours). The penultimate row denotes the maximal power consumption, ϱ (kW). The simulations in this article are based on a heterogeneous collection of loads in which the parameters for the TCLs take on values within these limits.

REFERENCES

- [1] *FPL On Call Saving Program*. Accessed: Feb. 2023. [Online]. Available: <http://www.fpl.com/save/programs/on-call.html>
- [2] M. Almassalkhi, L. D. Espinosa, P. D. Hines, J. Frolik, S. Paudyal, and M. Amini, “Asynchronous coordination of distributed energy resources with packetized energy management,” in *Energy Markets and Responsive Grids: Modeling, Control, and Optimization*, S. Meyn, T. Samad, I. Hiskens, and J. Stoustrup, Eds. New York, NY, USA: Springer, 2018, pp. 333–361.
- [3] M. Almassalkhi, J. Frolik, and P. Hines, “Packetized energy management: Asynchronous and anonymous coordination of thermostatically controlled loads,” in *Proc. Amer. Control Conf. (ACC)*, May 2017, pp. 1431–1437.
- [4] D. Angeli and P.-A. Kountouriotis, “A stochastic approach to ‘dynamic-demand’ refrigerator control,” *IEEE Trans. Control Syst. Technol.*, vol. 20, no. 3, pp. 581–592, May 2012.
- [5] G. Balas, R. Chiang, A. Packard, and M. Safonov, “Robust control toolbox user’s guide,” MathWorks, Natick, MA, USA, Tech. Rep., 2010.
- [6] P. Barooah, A. Buic, and S. Meyn, “Spectral decomposition of demand-side flexibility for reliable ancillary services in a smart grid,” in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 2700–2709.
- [7] A. Brooks, E. Lu, D. Reicher, C. Spirakis, and B. Wehl, “Demand dispatch,” *IEEE Power Energy Mag.*, vol. 8, no. 3, pp. 20–29, May 2010.
- [8] A. Busic and S. Meyn, “Distributed randomized control for demand dispatch,” in *Proc. IEEE 55th Conf. Decis. Control (CDC)*, Dec. 2016, pp. 6964–6971.
- [9] N. Cammardella, A. Busic, Y. Ji, and S. Meyn, “Kullback–Leibler–Quadratic optimal control of flexible power demand,” in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, Dec. 2019, pp. 4195–4201.
- [10] N. Cammardella, J. Mathias, M. Kiener, A. Busic, and S. Meyn, “Balancing California’s grid without batteries,” in *Proc. IEEE Conf. Decis. Control (CDC)*, Dec. 2018, pp. 7314–7321.
- [11] N. J. Cammardella, R. W. Moye, Y. Chen, and S. P. Meyn, “An energy storage cost comparison: Li-ion batteries vs distributed load control,” in *Proc. Clemson Univ. Power Syst. Conf. (PSC)*, Sep. 2018, pp. 1–6.
- [12] H. Chavez, R. Baldick, and S. Sharma, “Regulation adequacy analysis under high wind penetration scenarios in ERCOT nodal,” *IEEE Trans. Sustain. Energy*, vol. 3, no. 4, pp. 743–750, Oct. 2012.

- [13] Y. Chen, A. Busic, and S. P. Meyn, "State estimation for the individual and the population in mean field control with application to demand dispatch," *IEEE Trans. Autom. Control*, vol. 62, no. 3, pp. 1138–1149, Mar. 2017.
- [14] Y. Chen, M. U. Hashmi, J. Mathias, A. Bušić, and S. Meyn, "Distributed control design for balancing the grid using flexible loads," in *Energy Markets and Responsive Grids: Modeling, Control, and Optimization*, S. Meyn, T. Samad, I. Hiskens, and J. Stoustrup, Eds. New York, NY, USA: Springer, 2018, pp. 383–411.
- [15] M. Chertkov and V. Y. Chernyak, "Ensemble control of cycling energy loads: Markov Decision Approach," in *IMA Volume on the Control of Energy Markets and Grids*. Cham, Switzerland: Springer, 2018.
- [16] A. Coffman, A. Bušić, and P. Barooah, "A unified framework for coordination of thermostatically controlled loads," 2021, *arXiv:2108.05840*.
- [17] A. R. Coffman, N. Cammardella, P. Barooah, and S. Meyn, "Aggregate flexibility capacity of TCLs with cycling constraints," *IEEE Trans. Power Syst.*, vol. 38, no. 1, pp. 52–62, Jan. 2023.
- [18] K. Dodrill, "Demand dispatch—intelligent demand for a more efficient grid." U.S. Dept. Energy Office Electr. Del. Energy Rel., Washington, DC, USA, Tech. Rep. DOE/NETL- DE-FE0004001, 2011.
- [19] L. A. D. Espinosa and M. Almassalkhi, "A packetized energy management macromodel with quality of service guarantees for demand-side resources," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 3660–3670, Sep. 2020.
- [20] L. A. D. Espinosa, A. Khurram, and M. Almassalkhi, "Reference-tracking control policies for packetized coordination of heterogeneous DER populations," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 6, pp. 2427–2443, Nov. 2021.
- [21] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. New York, NY, USA: Wiley, 1986.
- [22] P. Fairley, "Energy storage: Power revolution," *Nature*, vol. 526, no. 7575, pp. S102–S104, Oct. 2015.
- [23] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Aggregate flexibility of thermostatically controlled loads," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 189–198, Jan. 2015.
- [24] A. Kasis, N. Monshizadeh, and I. Lestas, "Stability of primary frequency control with on-off load side participation in power networks," in *Proc. IEEE Conf. Decis. Control (CDC)*, Dec. 2018, pp. 6818–6823.
- [25] B. J. Kirby, "Frequency regulation basics and trends," Rep. Prepared U.S. Oakridge Nat. Lab., Oak Ridge, TN, USA, Tech. Rep. DoE ORNL/TM-2004/291, 2004.
- [26] S. Koch, J. L. Mathieu, and D. S. Callaway, "Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services," in *Proc. 17th Power Syst. Comput. Conf.*, 2011, pp. 1–7.
- [27] P. Kundur, *Power System Stability and Control, of EPRI Power System Engineering*, vol. 7. New York, NY, USA: McGraw-Hill, 1994.
- [28] H. Kwakernaak, "Robust control and H_∞ -optimization—Tutorial paper," *Automatica*, vol. 29, no. 2, pp. 255–273, Mar. 1993.
- [29] R. Malhame and C. Chong, "Electric load model synthesis by diffusion approximation of a high-order hybrid-state stochastic system," *IEEE Trans. Autom. Control*, vol. AC-30, no. 9, pp. 854–860, Sep. 1985.
- [30] J. Mathias, "Balancing the power grid with distributed control of flexible loads," Ph.D. thesis, Dept. Elect. Comput. Eng., Univ. Florida, Gainesville, FL, USA, 2022.
- [31] J. Mathias, A. Busic, and S. Meyn, "Demand dispatch with heterogeneous intelligent loads," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 1–10.
- [32] J. Mathias, R. Kaddah, A. Buic, and S. Meyn, "Smart fridge/ dumb grid? Demand dispatch for the power grid of 2020," in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 2498–2507.
- [33] J. Mathieu, M. Dyson, D. Callaway, and A. Rosenfeld, "Using residential electric loads for fast demand response: The potential resource and revenues, the costs, and policy recommendations," *Proc. ACEEE Summer Study Buildings*, Pacific Grove, CA, USA, 2012, pp. 189–203.
- [34] J. L. Mathieu, S. Koch, and D. S. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 430–440, Feb. 2013.
- [35] S. P. Meyn, P. Barooah, A. Busic, Y. Chen, and J. Ehren, "Ancillary service to the grid using intelligent deferrable loads," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2847–2862, Nov. 2015.
- [36] S. Meyn, P. Barooah, A. Bušić, and J. Ehren, "Ancillary service to the grid from deferrable loads: The case for intelligent pool pumps in Florida," in *Proc. Conf. Dec. Control*, Dec. 2013, pp. 6946–6953.
- [37] V. S. K. Murthy Balijepalli, V. Pradhan, S. A. Khaparde, and R. M. Shereef, "Review of demand response under smart grid paradigm," in *Proc. ISGT-India*, Dec. 2011, pp. 236–243.
- [38] B. M. Sanandaji, H. Hao, K. Poolla, and T. L. Vincent, "Improved battery models of an aggregation of thermostatically controlled loads for frequency regulation," in *Proc. Amer. Control Conf.*, Jun. 2014, pp. 38–45.
- [39] R. Smith, K. Meng, Z. Dong, and R. Simpson, "Demand response: A strategy to address residential air-conditioning peak load in Australia," *J. Modern Power Syst. Clean Energy*, vol. 1, no. 3, pp. 223–230, Dec. 2013.
- [40] S. H. Tindemans, V. Trovato, and G. Strbac, "Decentralized control of thermostatic loads for flexible demand response," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 5, pp. 1685–1700, Sep. 2015.
- [41] E. Vrettos, C. Ziras, and G. Andersson, "Fast and reliable primary frequency reserves from refrigerators with decentralized stochastic control," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2924–2941, Jul. 2017.
- [42] A. J. Wood, B. F. Wollenberg, and G. B. Sheblé, *Power Generation, Operation, and Control*, 3rd ed. New York, NY, USA: Wiley, 2013.
- [43] C. Zhao, U. Topcu, N. Li, and S. Low, "Design and stability of load-side primary frequency control in power systems," *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1177–1189, May 2014.
- [44] C. Ziras, S. You, H. W. Bindner, and E. Vrettos, "A new method for handling lockout constraints on controlled TCL aggregations," in *Proc. Power Syst. Comput. Conf. (PSCC)*, Jun. 2018, pp. 1–7.



Joel Mathias received the B.Eng. degree in electronics and communications engineering from the University of Mumbai, Maharashtra, India, in 2010, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2014 and 2022, respectively.

He was a Research Associate with the Tata Institute of Fundamental Research, Mumbai, India, from 2011 to 2012, and his industry experience includes Electric Power Engineers, Inc., Austin, TX, USA, and Tata Consultancy Services, Mumbai. For his graduate studies with the USA, he has received the J.N. Tata Endowment for the higher education of Indians, the Lady Navajbai Tata Scholarship, and achievement awards from the University of Florida. He is currently a Post-Doctoral Fellow with Arizona State University, Tempe, AZ, USA. His research interests are in stochastic processes, control theory, and optimization, with applications in the area of smart power grids.



Ana Bušić (Member, IEEE) received the M.S. degree in mathematics and the Ph.D. degree in computer science from the University of Versailles, Versailles, France, in 2003 and 2007, respectively.

She was a Post-Doctoral Fellow with Inria Grenoble Rhone-Alpes and the University Paris Diderot-Paris 7, Paris, France. She is currently a Research Scientist with Inria Paris and the Computer Science Department, École Normale Supérieure, PSL University, CNRS, Paris. She is also a member of the Laboratory of Information, Networking and Communication Sciences (LINCS). Her research interests include stochastic modeling, reinforcement learning, simulation, and performance evaluation, with applications to networks and power systems.

Dr. Bušić was a recipient of the 2015 Google Faculty Research Award.



Sean Meyn (Fellow, IEEE) received the B.A. degree in mathematics from the University of California, Los Angeles (UCLA), Los Angeles, CA, USA, in 1982. He is currently pursuing the Ph.D. degree with Peter Caines, McGill University, Montreal, Canada.

After about 20 years as a Professor of ECE with the University of Illinois, Urbana-Champaign, in 2012, he moved to beautiful Gainesville. He is currently a Professor and a Robert C. Pittman Eminent Scholar Chair with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA, the Director of the Laboratory for Cognition and Control, and Inria International Chair with INRIA, Paris, France. He is also an IEEE CSS Distinguished Lecturer. His interests span many aspects of stochastic control, stochastic processes, information theory, and optimization. For the past decade, his applied research has focused on engineering, markets, and policy in energy systems.