Spectral Decomposition of Demand-Side Flexibility for Reliable Ancillary Services in a Smart Grid

Prabir Barooah
Univ. of Florida
pbarooah@ufl.edu

Ana Bušić
Inria/ENS
ana.busic@inria.fr

Sean Meyn
Univ. of Florida
meyn@ece.ufl.edu

Abstract *This paper describes a new way of thinking about demand-side resources to provide ancillary services to control the grid. It is shown that loads can be classified based on the frequency bandwidth of ancillary service that they can offer. If demand response from loads respects these frequency limitations, it is possible to obtain highly reliable ancillary service to the grid, while maintaining strict bounds on the quality of service (QoS) delivered by each load. It is argued that automated demand response is required for reliable control. Moreover, some intelligence is needed at demand response loads so that the aggregate will be reliable and controllable.*

1 Introduction

A key engineering challenge in integrating large amounts of renewable energy into the electric grid is uncertainty and volatility of renewable sources. Fig. 1 shows evidence of the large intermittency of renewable energy in the U.S. Pacific Northwest. The grid operators in this region are currently searching for additional controllable resources to combat this volatility.

Management of the grid today can be cast as a massive disturbance rejection problem. Generation and demand of electric power must be balanced at all scales to ensure stable grid operation. This is achieved by sensor measurements and manipulation of controllable resources such as gas-turbine generators to regulate voltage, frequency and phase to constant values. Responsive generators supply not just power, but *ancillary services* — resources to regulate the grid.

As greater renewable penetration brings more volatility to the grid, this paradigm of controllable generation matching demand is no longer tenable.

Many BAs employ demand response (DR) programs that use controllable loads to reduce peak demand and manage emergency situations [22]. Florida Power and Light (FPL), for example, has 780,000 residential customers enrolled in their *OnCall Savings Program* which allows FPL to remotely turn off select equipment - such as pool pumps - when needed [1]. Today, FPL uses this service only 3–4 times per year.

Increasingly, loads are providing non-emergency ancillary service as well. Alcoa provides 70 MW of frequency regulation service to the mid-continent ISO (MISO) by providing control over their aluminum smelting operation in Indiana [26]. Growth of these resources in these wholesale markets has helped lower costs per megawatt-hour from 2009 to 2011 [20].

1.1 Virtual Storage from Flexible Loads

It is often said that renewable energy is expensive because of the high cost of grid-level energy storage. In the absence of large, expensive batteries, we may have to increase our inventory of responsive fossil-fuel generators, negating the environmental benefits of renewable energy.

We counter that *storage is everywhere*: We do not need to rely entirely on expensive batteries or fast-responding fossil fuel generators to track zero-mean regulation signals or balancing reserves.

There is enormous flexibility in the power consumption of the majority of electric loads. This flexibility can be exploited to create "virtual batteries". The best example of this is the heating, ventilation, and air conditioning (HVAC) system of a building: There is no perceptible change to the indoor climate if the airflow rate is increased by 10% for 20 minutes, and decreased by 10% for the next 20 minutes. Power consumption deviations follow the airflow deviations closely, but indoor temperature will be essentially constant.

In this paper *capacity* is always used in terms of power rather than energy. We will show that the capacity of ancillary service from a class of loads is a function of timescale, or equivalently, frequency of ancillary service.

The potential capacity of buildings in the U.S. is enormous since HVAC systems account for greater than 30% of the electricity consumed in the United States, and buildings overall account for 74% of overall consumption [28]. Moreover, their collective thermal inertia is large, so the potential for virtual storage is enormous. For example, the capacity from fans in commercial build-

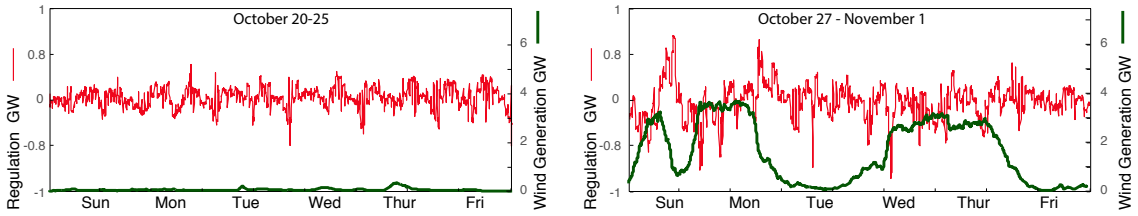


Figure 1: Volatility due to renewables is increasing in the Pacific Northwest: Balancing reserves from two typical weeks in the Fall of 2013.

ing HVAC systems in the U.S. is approximately 6 GW in the time of scale of a few seconds to a few minutes [12, 12]. This flexibility is much greater at lower frequencies: Capacity is estimated to be 42 GW on timescales ranging from ten minutes to an hour [16]. A unified approach to estimating capacity as a function of frequency is presented in Section 2.2 of the present paper.

The thermal inertia of the building is analogous to a battery that is charged and discharged to supply ancillary service to the grid. This analogy is illustrated in Fig. 2, which shows how the power consumption deviation from a baseline value can be manipulated exactly as a battery can be charged and discharged. Both batteries and loads can provide ancillary service in the form of power trajectories that are *zero energy* on average, which is a feature of regulation signals such as those shown in Fig. 1.

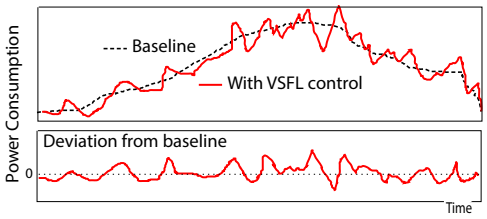


Figure 2: Virtual Storage from Flexible Loads.

To manipulate the power consumption around the baseline value, while maintaining quality of service (QoS) to the consumer, the loads must be equipped with 1) local intelligence, and 2) information regarding the state of the grid.

Local intelligence is already present in modern commercial buildings — only a software add-on is required to execute additional intelligence for virtual storage. Consequently, virtual storage can be obtained at very low cost.

The acronym VSFL that is used throughout the paper is short-hand for *Virtual Storage from Flexible Loads*. This represents a grid level control architecture in which loads act in parallel with traditional controllable generation to provide ancillary service according to their capacity and bandwidth constraints. Control design respects the needs of both the grid and the consumer. In particular the function served by a load cannot be compromised. VSFL is not *direct load control* — there is no central agent who is selectively interfering with loads.

Fig. 3 illustrates the foundation of the decentralized control architecture proposed in this work. It is not very different from today's architecture in which controllable generators act as *actuators* in a grid-level control system. An example of the signal that generators routinely track today is the Area Control Error (ACE) that is broadcast by the BA. In the future, generators will be replaced by "intelligent loads" that will provide cheaper and more reliable ancillary service. Intelligence at each load is required so that it can provide reliable ancillary service to the grid, while maintaining QoS to consumers.

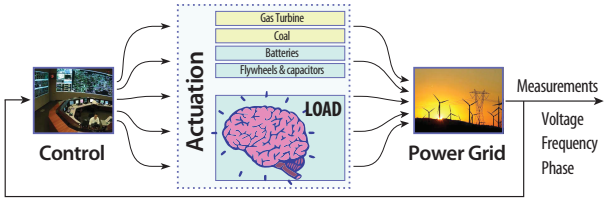


Figure 3: Power grid as a feedback control system.

1.2 Classification of resources

A focus of this paper is the *actuation block* in Fig. 3. The collection of resources in the actuator block will continue to grow, and making the best use of them is the control problem we must solve. These resources are differentiated by several factors, including,

(i) Bandwidth of regulation. Coal generators can ramp up and down slowly to address low frequency volatilities, while other resources can take care of the high frequency component of the control signal.

(ii) Dynamics. BAs have rules on maximal delay and other grid-level quality of service metrics.

(iii) Costs. These include installation, start-up, running cost, and pollution.

The wide-band balancing reserves shown in Fig. 1 could be obtained from many resources operating in parallel, even if each resource can provide only a limited portion of the total bandwidth. By restricting the bandwidth of service it is possible to deliver reliable ancillary

service to the grid at reduced cost. In the case of ancillary service obtained from flexible loads, bandwidth constraints also help to enforce strict bounds on the quality of service (QoS) offered by the load to the consumer.

Case studies are provided in Section 3 to show how this can be accomplished with little additional cost, through the use of appropriate control architectures and algorithms.

1.3 Evaluating VSFL

What are the indicators of success for ancillary service based on flexible loads? We claim that the following issues must be addressed if a demand-based ancillary service is to be called "successful":

(i) *Is the service of high quality?* Does the deviation in power consumption accurately track the desired deviation target?

(ii) *Is the service reliable?* Does the grid operator have assurance that ancillary service will be available each day? Ancillary service capacity may vary with time, but capacity must be predictable.

(iii) *Is it cost effective?* This includes installation cost, communication cost, and maintenance.

(iv) *Is the incentive to the consumer reliable?* If a consumer receives a $50 payment for one month, and only $1 the next, will there be an explanation that is clear to the consumer?

(v) *Does the system meet QoS constraints?* A pool must be clean, building climate is subject to strict bounds, farm irrigation is subject to strict constraints, and data centers require sufficient power to perform their tasks.

The first item is a long-standing concern of grid operators. Ancillary service from some generators can in fact be harmful to the grid because of significant phase lag in generation response [11, 14]. In many markets ancillary service providers must pass tests to prove that they can follow a regulation signal within a specified response time [24, 25].

There may also be high cost to generators that ramp up and down to provide ancillary service, as well as opportunity cost because the generator cannot run at capacity [3].

These five criteria can be met using the VSFL framework described in this article. Automated demand response can be designed to ensure that strict bounds on QoS are guaranteed. Automation and attention to QoS constraints also addresses common concerns in applying demand response: demand rebound, consumer fatigue, and free-rider behavior [7]. These three potential risks are eliminated through design.

Literature review The use of flexible loads to provide ancillary service via demand response that is *automated* is considered in many recent papers. Loads that have been considered include commercial building HVAC fans in the time-scale of seconds to nearly one hour [10, 13, 18], thermostatic devices that can provide ancillary service in the time-scale of a few minutes [4, 21, 22] (and refs. therein), electric vehicle charging that can provide ancillary service in the time scale of a few hours [9, 19, 21, 27], and pool pumps in the states of Florida or California can provide ancillary service on longer time scales [8, 23] (these loads are also used for peak-shaving [1, 5]).

Two papers presented at the 2013 HICSS meeting considered frequency domain aspects of demand response. In [6] the authors point to previous studies on the power spectral density of wind, arguing that much of the volatility is concentrated in timescales of an hour or more. The article [7] focuses on the inherent periodicity of loads, and how this represents one constraint in obtaining demand response at low frequencies. Both of these papers focus on real time prices rather than automated demand response.

The remainder of this paper is organized as follows. Section 2 provides more details on the needs of the grid in terms of infrastructure and control. Section 3 concerns harnessing the flexibility of aggregates of deferrable loads, and Section 4 contains conclusions and topics for future research.

Acknowledgements The experimental work described in this paper would not be possible without help from many partners. The physical plants division gave access to Pugh Hall at the University of Florida for experiments on the HVAC system. We thank Tim Middelkoop for developing the software to make these experiments feasible.

This paper is based on collaboration with students and other colleagues. In particular, Yashen Lin conducted the experiments at Pugh hall that is summarized in Section 3.1, and the numerical experiments summarized in Section 3.2 were conducted by Yue Chen. Samuel Putnam gave many suggestions for improving the presentation.

Research supported by the NSF grants CPS-0931416 and CPS-1259040, the French National Research Agency grant ANR-12-MONU-0019, and US-Israel BSF Grant 2011506.

2 Temporal Classification

The frequency-based classification of resources proposed here is a refinement of what is done today: An-

cillary services from online generation are organized today in terms of the time-scale of regulation provided. In one classification, there is *Primary Control, Secondary Control, and Tertiary Control*. In another, they are categorized as *Regulation* (seconds to minutes), *Load Following* (minutes to hours), and *Unit Commitment* (hours to days).

The existing taxonomy of ancillary services is inadequate in a world with significant volatile generation. The downward ramps in wind generation seen in Fig. 1 appear similar to outages of large generators. Hence the separation between contingency and normal operation will blur with highly variable generation: Load following will become *renewable following*, and automatic generation control (AGC) will no longer be relevant to generators alone.

Our proposal is a more fine-grain classification of resources based on the bandwidth of service that they can offer, and also the reliability of the resource.

Batteries can provide a large bandwidth of ancillary service, with a potentially higher cost in terms of battery life at higher frequency ranges [15]. Heating and ventilation systems in commercial buildings can provide service in a high frequency band, corresponding to periods ranging from a few seconds to one hour [13,16]. The capacity from commercial buildings is enormous, so that today more than half of the regulation needs in the U.S. could be provided from this source alone.

A large collection of on/off loads such as chillers, residential air-conditioners, or pool pumps, can provide ancillary service at lower frequencies, on the order of several hours to one day. At the lowest frequencies of a day or more, an important resource will be flexible manufacturing, in which production will ramp up and down depending on the availability of energy from the sun or wind.

We propose that the BA decompose a regulation signal into various components based on frequency, and resources can serve these components. A resource may be a generator, a battery, or a collection of similar flexible loads. Each resource may be classified by a pair of power spectral densities (S,T) where S is the power spectral density of ancillary service that can be provided, and T is the power spectral density of tracking error. Each resource may be capable of providing many different values for S and T. Spectral constraints on ancillary service are illustrated in the following examples.

2.1 On/off loads

Consider a large number of on/off loads, such as pool pumps. In the state of Florida there are approximately one million pools, each using a pool pump that consumes approximately one kW of power when in operation. These are on/off loads that typically perform just one cleaning cycle per day, running from 6 to 12 hours [5].

In the case of pools, QoS constraints include bounds on the cleaning period each day and each week, and also the number of times the pump is turned on and off. We focus on the second constraint because of the interesting translation to constraints on the power spectral density of ancillary service.

In a model to approximate potential bandwidth of ancillary service we let N denote the number of pools, and $g \approx 1kW$ denote the power consumption of one pool pump.

Consider first this simple case: Suppose that we would like to track a square wave over a period of 24 hours. We impose a hard QoS constraint: each pool can switch on or off exactly once each day. Let M denote the number of times the square wave changes sign, which is twice the number of cycles. Hence the frequency in cycles/hour is $F = M/48$. To obtain bounds on capacity as a function of frequency requires additional modeling of constraints on the loads.

To track a single square wave with magnitude A (kW), assume initially that half are on and half are off. Divide the loads into N/M classes, each of which switches exactly once. Assume that the loads in a single class have the same on/off state. Hence, when the loads in a class switch, they all do so in the same direction (all off, or all on). The change in power when one class switches is $\pm gN/M$, which corresponds to a square wave with one half this magnitude (taking values $\pm\frac{1}{2}gN/M$). This gives a bound on capacity that decreases linearly with frequency:

$$A \le \tfrac{1}{2}\frac{gN}{M} = \frac{gN}{96F} \qquad \text{(kW)}$$

The extension to more general signals can be obtained via a spectral decomposition, along with another decomposition of the total collection of N pools into 'frequency classes'. Consider a decomposition of the N loads into m "f-classes". The size of the kth class is q_kN, where $\sum q_k \le 1$.

Let $\{v_k(t) : 1 \le k \le m\}$ denote square waves with unit amplitude, and respective frequencies $\{F_k\}$. The previous bound implies that the kth class can track A_kv_k provided $|A_k| \le q_kgN/96F_k$. This holds for each k and some sequence $\{q_k\}$ if and only if

$$\sum_k F_k|A_k| \le gN/96.$$

Under this condition, the collection of N loads can track $v(t) = \sum A_kv_k(t)$.

Any square-integrable function on $[0,24]$ can be represented as an infinite sum, $v(t) = \sum_{k=0}^{\infty} A_kv_k^*(t)$, $0 \le t \le$

24, where the convergence is in L_2. The functions in this representation form an orthonormal basis: $v_0^*(t) \equiv 1$, and v_k^* is the square wave with exactly k cycles on the time-interval $[0,24]$ when $k \geq 1$. In this case $F_k = k/24$ cycles/hour, which gives the bound,

$$\sum_{k=1}^{\infty} k|A_k| \leq gN/4$$

These calculations show how a single constraint on QoS (the number of state changes in one day) can be translated to a frequency dependent bound on capacity. The frequency constraint will become less restrictive if the consumers allow a greater number of state changes each day.

The ℓ_1-norm bound on the Fourier coefficients $\{A_k\}$ will impose constraints on the possible power spectral densities S that can be delivered by this class of loads. We believe that the tracking error power spectral density T will be nearly zero with appropriate design [23].

2.2 Continuously variable thermal loads

Approximately 30% of the U.S. commercial building floor space is equipped with Variable Air Volume (VAV) systems in which the HVAC loads can be varied continuously [2]. Capacity-bandwidth constraints for these buildings can be derived based on standard linear dynamic models. The QoS constraints in this case include temperature, humidity, and weekly operating cost.

Consider a single building with an air-conditioning unit whose power consumption can be varied continuously. The dynamics of indoor temperature T in the building can be modeled by a resistor-capacitor (RC) network model. Here we use a first-order model:

$$C\tfrac{d}{dt}T(t) = -\frac{1}{R}(T(t) - T_A(t)) + Q_{\text{gs}}(t) - Q_{\text{A/C}}(t)$$

in which C (J/K) is the total thermal capacity of the building, R (K/W) is the effective thermal resistance of the building's envelope to heat exchange between the interior space and the ambient, T_A is the ambient temperature, Q_{gs} (W) is the total rate of heat generated inside the building plus solar heat gain, and $Q_{\text{A/C}}$ (W) is the rate of cooling provided by the A/C (air conditioner).

The electric power consumption P is approximated by the cooling power provided, $Q_{\text{A/C}}$. This is only an approximation since there is efficiency loss, and phase lag between power consumed and cooling power provided. The approximation $P = Q_{\text{A/C}}$ is taken for granted in the calculations that follow.

The baseline behavior of the building is taken to be static. Using an asterix "*" to denote steady-state quantities, the steady state cooling Q^* provided by the A/C and the steady state indoor temperature T^* satisfy the linear equation,

$$0 = -\frac{1}{R}(T^* - T_A^*) + Q_{\text{gs}}^* - Q_{\text{A/C}}^*$$

The deviation about these steady state values satisfy,

$$\tfrac{d}{dt}y(t) = -\frac{1}{\tau}y(t) - \frac{1}{C}v(t),$$

in which $y(t) = T(t) - T^*$, $v(t) = Q_{\text{A/C}}(t) - Q_{\text{A/C}}^*$, and $\tau = RC$ is the *time constant* of this ODE. The transfer function from cooling power deviation v to indoor temperature deviation y is

$$H(s) = \frac{R/\tau}{s + 1/\tau}$$

Suppose that v is required to track the sinusoid with magnitude A and frequency ω, while maintaining a strict bound on indoor temperature, $|y(t)| \leq \Delta_T$ for each t. If this QoS requirement is only imposed in steady-state, then it can be re-interpreted as a bound on the transfer function, $|H(j\omega)| \leq \Delta_T/A$. Equivalently,

$$\omega^2 \geq (\frac{A}{\Delta_T C})^2 - \frac{1}{\tau^2} = \frac{1}{C^2}(\frac{A^2}{\Delta_T^2} - \frac{1}{R^2})$$

Expressing A as a fraction of steady-state power consumption, $A = \delta Q_{\text{A/C}}^*$, with $0 < \delta < 1$, we arrive at a single bound on frequency as a function of QoS and building parameters:

$$\omega^2 \geq \frac{1}{C^2}\frac{\delta^2 (Q_{\text{A/C}}^*)^2}{\Delta_T^2} - \frac{1}{R^2} = \frac{\delta^2}{\Delta_T^2}\frac{(Q_{\text{A/C}}^*)^2}{C^2} - \frac{1}{\tau^2} \qquad (1)$$

Upper bounds on ω can be obtained based on characteristics of the building hardware rather than building dynamics. In most cases there will be a cost in terms of wear and tear that increases with frequency.

Extension to more general signals can be obtained via a spectral decomposition, exactly as in Section 2.1.

Commercial buildings typically have large values of R and C when compared with residential homes, and hence a large time-constant τ. This makes them attractive for VSFL.

There is however tremendous potential value from residential homes because there are so many of them. To understand the bandwidth of service that can be obtained, consider a typical home in which $Q_{\text{A/C}}^* = 7000$ W. We take $\delta = 0.1$, so that this typical home will be asked to vary its consumption by ± 700 W. For a city with one million residential homes of this size, the grid will receive 700 MW of capacity.

Most homeowners in the U.S. today will accept temperature deviations of one degree Fahrenheit, which is approximately $\Delta_T = 0.5$ degrees Kelvin.

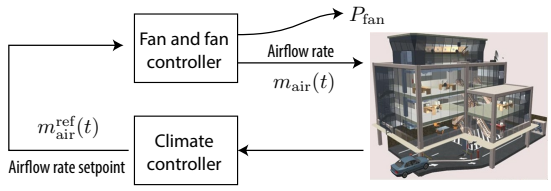


Figure 4: Climate control loop in a VAV HVAC system.

The range of possible values for R and C is large, depending on climate and age. Consider the following range of values: $C \in [5 \times 10^6,\ 50 \times 10^6]$ (J/K) and $R \in [1/10^4,\ 1/500]$ (K/W). The time constant τ then varies from 0.14 hours to 27 hours.

Given this range of time constants and other parameters, the lower bound (1) takes values in the range $[0.671,\ 6.4] \times 10^{-4}$ rad/s. That is, under these conditions it is possible to extract the specified flexibility for time-scales bounded above by 2.5 hours to 25 hours at 700 MW of capacity for a city with one million homes.

Presently, most residential HVAC systems are on/off rather than continuously variable. Methods to obtain ancillary service from these loads are described in Section 3.2.

3 Distributed Intelligence for Reliable Ancillary Services

We now consider how to harness many different sources of generation and millions of flexible loads to create a reliable actuator block for control of the grid.

Just as a generator adjusts its power output today based on a regulation signal such as shown in Fig. 1, flexible loads will respond to a signal from the BA. In response, each load will adjust its consumption up and down – perhaps by just a small percentage of its average power consumption.

In the hierarchical architecture proposed, intelligence resides at both the grid-layer and node-layer, while in centralized or distributed architectures, intelligence at one layer is absent. Intelligence at each load is required so that it can can provide ancillary service to the grid, while maintaining QoS constraints.

The manifestation of "intelligence" will depend on the physical properties of the load as well as the QoS constraints. We illustrate the synthesis of local control algorithms in two examples contained in the next two subsections.

3.1 Continuously variable loads

The approach of [12] summarized here illustrates how to design VSFL to achieve the criteria for success summarized in Section 1.3.

Consider a large commercial building equipped with a VAV HVAC system. The building receives a reference signal ΔP_{BA} from the BA just as a generator does today. A bandpass filter generates the local reference signal, denoted ΔP_{ref}. Each fan in the building will track a scaled version of this signal.

A fan controller within the building climate control system varies the fan speed to maintain the airflow rate at the setpoint commanded by the climate controller (see Fig. 4). Measurements of the power consumed by the motor are available to the climate control system.

By adding a command to the baseline airflow rate setpoint $\frac{d}{dt} m^{\text{ref}}_{\text{air}}$ computed by the building's climate controller, the fan speed can be increased or decreased around the baseline. Fig. 5(a) shows a schematic of the proposed local intelligence at the building for VSFL. The command u_{VSFL} is suitably designed by the feedback controller C so that the power consumption deviation from the baseline tracks the reference ΔP_{ref} in real time.

The design of the bandpass filter that defines ΔP_{ref} is based on the physical characteristics of the building, as well as the building's climate control system. A high pass filter is chosen for three distinct reasons:

(i) The large thermal inertia of the building acts as a low pass filter to fast deviations in the airflow rate. Hence a high pass filter ensures that indoor climate hardly changes.

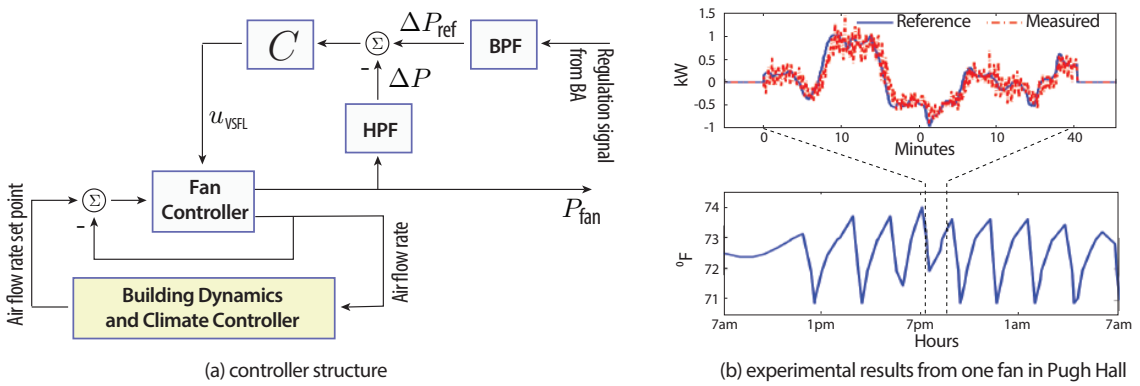
(ii) The existing climate control system in the building that regulates temperature is also a low-pass filter (typically PI). This means that the climate control system will not try to reject the command u_{VSFL} that enters the fan control loop as a disturbance.

(iii) The baseline power is defined to be the power that would have been consumed if the VSFL controller were absent. Based on a separation of time-scales, baseline power can be estimated using another high-pass-filter (HPF in Fig. 5(a)).

The method has been experimentally demonstrated in Pugh Hall in the UF campus [17]. Fig. 5(b) shows tracking results from the real-time feedback control experiment. The regulation signal RegD from PJM (an ISO: www.pjm.com) was used as the BA reference ΔP_{BA}. The controller C_{VSFL} was a lag compensator, designed based on the experimentally identified transfer function from airflow rate to fan power.

Analysis of the tracking error included in [17] shows that the system meets the criteria set by PJM for a resource to take part in its ancillary service market.

The controller providing VSFL does not cause any noticeable change in the indoor climate of the building. See Fig. 5(b): there is no statistical difference between the variations in the temperature during the 40 min period in



(a) controller structure

(b) experimental results from one fan in Pugh Hall

Figure 5: VSFL from a single HVAC fan in a commercial building.

which the test took place and the rest of the day.

3.2 Ancillary services from on/off loads

Control of on-off loads loads requires very different approaches compared to the variable speed HVAC systems treated in the previous section. We focus on residential pool pumps to simplify discussion, following the discussion in Section 2.1. The range of application of the techniques described here are far broader.

The control architecture proposed here is designed to address privacy concerns and communication constraints. To provide ancillary service in a specified frequency band, we argue that it is essential to introduce randomization. This avoids synchronization, much like randomized congestion avoidance protocols in communication networks.

In the commercial building application considered previously, communication was assumed in one direction only: From grid to building. Here it is assumed that the BA has real-time estimates of the power consumed by the collection of pools, denoted $\{\gamma_t\}$. The BA will send a common signal denoted $\{\zeta_t\}$ to each load that will depend on these measurements and the desired reference signal $\{r_t\}$. The BA will use *actuator control* to shape the power consumption of the aggregate. This will take the form,

$$\zeta_t = G_a e_t \tag{2}$$

where $e_t = r_t - \gamma_t$, and G_a is the control algorithm (e.g., a PI controller).

The construction of a randomized policy for an individual pool pump begins with a Markovian model for the load. We consider a discrete-time, finite state-space model. The state for an individual pool is denoted X, which evolves on the state space

$$\mathsf{X} = \{(m,i) : m \in \{\oplus,\ominus\},\ i \in \{1,\ldots,T\}\};$$

$X(t) = (\ominus, i)$ indicates that the pool-pump is currently off and has remained off for exactly i time units, and $X(t) = (\oplus, i)$ represents the alternative that the pool-pump has been operating continuously for exactly i time units.

The randomized algorithm is based on a family of transition matrices $\{P_\zeta : \zeta \in \mathbb{R}\}$. The pool system will change its state according to the controlled Markovian dynamics,

$$\mathsf{P}\{X(t+1) = y \mid X(t) = x\} = P_{\zeta_t}(x,y)$$

where $\{\zeta_t\}$ denotes the signal broadcast by a BA. In current research it is assumed that the BA measures the aggregate power consumption of the collection of pools, and uses this data (along with other grid measurements) to generate the signal ζ.

A model of the aggregate power consumption is valuable for design of the randomized policy and also the control algorithm used at the BA that determines ζ. This model is obtained as an infinite-N limit, with N equal to the number of pools.

On letting $X^i(t)$ denote the state of the ith pool, the following limit will hold under mild assumptions:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{X^i(t) = x\} = \mu_t(x), \quad x \in \mathsf{X}.$$

With μ_t treated as a row vector, and P_{ζ_t} as a matrix, the sequence of probability measures can be regarded as the state in a controlled nonlinear state space model,

$$\mu_{t+1} = \mu_t P_{\zeta_t} \tag{3}$$

which is linear in the "state" μ_t, and necessarily nonlinear in the "control" ζ_t. The output of this model will be denoted γ_t, and taken to be equal to the number of pools that are operating.

There are several questions to answer:

(i) How to choose the family of transition matrices?

(ii) The system behavior is inherently nonlinear – how can we be sure that this will provide reliable ancillary service?

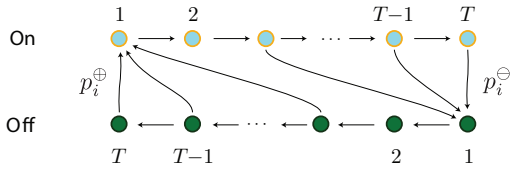


Figure 6: State transition diagram for the pool-pump model.

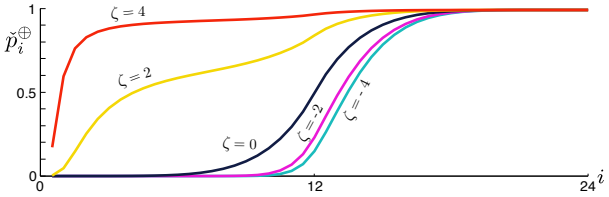


Figure 7: Probability vectors $p_\zeta^\oplus$ and $p_\zeta^\ominus$ define the randomized policy P_ζ.

(iii) Can QoS constraints be guaranteed to each consumer?

The first two topics are treated in [23]. The transition matrix P_ζ was constructed as the optimal solution to a certain stochastic optimal control problem. Topic (iii) is addressed in [8].

A linearization about an equilibrium can provide insight into the behavior of the nonlinear model (3). Suppose that the control-free model has a unique equilibrium, denoted π. This is simply an invariant probability measure for P_0, which is assumed to be unique. Because we are assuming that each pool will change its behavior only slightly, it is reasonable to assume that $\mu_t \approx \pi$ for all t, and we also assume that $\zeta_t \approx 0$ to obtain a linearization.

The state dynamics can be approximated as follows when $\pi \approx \mu_t$:

$$\mu_{t+1} = \mu_t[P_{\zeta_t} - P_0] + \mu_t P_0 \approx \pi[P_{\zeta_t} - P_0] + \mu_t P_0$$

We assume that P_ζ is differentiable at $\zeta = 0$ to justify the Taylor series approximation $\mu_{t+1} \approx \mu_t P_0 + B^\intercal \zeta_t$, where $B^\intercal = \frac{d}{d\zeta}\pi P_\zeta$, evaluated at $\zeta = 0$. This leads to the linear state space model,

$$\Phi_{t+1} = A\Phi_t + B\zeta_t, \qquad y_t = C\Phi_t \tag{4}$$

in which Φ_t approximates the deviation $(\mu_t - \pi)^\intercal \in \mathbb{R}^{2T}$, $A = P_0^\intercal$, and the row vector C is defined so that,

$$y_t = \sum_{i=1}^{T} \Phi_t(\oplus, i)$$

Hence y_t is an approximation of γ_t.

To show how well the linearization predicts performance of the nonlinear stochastic system, we turn to the numerical example of [23].

A state-transition diagram for the nominal model is shown in Fig. 6. The time T is an upper bound on the number of time units the pool pump can stay on or off. The pool switches from on to off (or off to on) with probability depending on the time it has been on (or off).

The controlled transition matrix for a load is defined by the values,

$$\begin{aligned} p_\zeta^\oplus(i) &:= P(\text{switch on} \,|\, \text{off } i \text{ hours}) \\ p_\zeta^\ominus(i) &:= P(\text{switch off} \,|\, \text{on } i \text{ hours}) \end{aligned} \tag{5}$$

Fig. 7 contains plots of the vectors $\{p_\zeta^\oplus\}$ for values $\zeta = 0, \pm 4, \pm 6$ that are used in one set of experiments in [23]. In this example $p_0^\oplus(i)$ is nearly deterministic (values near zero or one), which is similar to the usual operation of the pool with a twelve-hour cleaning cycle.

Results from two experiments are shown in Fig. 8. A typical regulation signal from the BPA is used (similar to those shown in Fig. 1). This is scaled and filtered using a low-pass filter. The signal ζ was obtained using PI control based on the actuator feedback control (2).

The first plot uses the controlled matrix with entries plotted in Fig. 7, and the second uses a similar controlled transition matrix based on an eight-hour nominal cleaning cycle. The tracking performance is remarkable in each case. In particular, it is surprising that a ± 400 MW signal can be tracked, given that the average power consumption of the pools is 500 MW in Scenario 1. There is no reason to believe that a network of one million pools cannot supply much more ancillary service to the grid.

4 Conclusions

The transformation of the power grid will require distributed resources and control. This paper has focused on approaches to harnessing load based resources, and has left out several important issues. We discuss three broad avenues for research in the remainder of this section.

4.1 Individual risk

Each class of loads has its own QoS constraints and costs. In the case of aluminum smelting, "... a reduction in process efficiency and stability, as well as increased maintenance costs are risks to supplying regulation services." Alcoa has been providing regulation service to MISO since 2009, which is an indication that these risks proved to be manageable [26].

The analysis in Section 3.2 has focused on service to the grid. What is the experience of an individual pool owner who agrees to participate in this scheme? An analysis of quality of service for an individual load can be performed by an extension of the mean field model. It should not be surprising that with 100,000 or more loads,

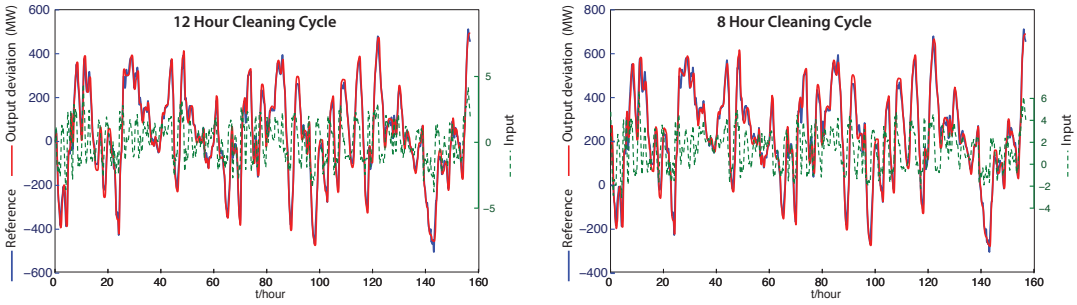


Figure 8: Closed loop simulation of 100,000 pools.

it is likely that at least one load will sometimes receive very bad service with a randomized policy, unless there is another local layer of control.

One approach is to add an additional counter at each pool that keeps track of the number of hours of cleaning over, say, the past 48 hours. During a time period for which this falls outside of a priori bounds, the pool will ignore the grid signal. The zero mean regulation signal is designed so that this is an uncommon event – at any given time, with very high probability, at most 10% of pools are out of range.

In preliminary numerical experiments we have seen a very small reduction in capacity, while maintaining the accurate tracking observed in Figure 8, and also *guaranteeing* strict bounds on QoS for each pool [8].

4.2 Standards for grid-friendly loads

A topic of future research is the formulation of standards for grid friendly appliances. If there is no impact on QoS and if the value to the grid is sufficiently large, then we might mandate that some HVAC systems be built with the communication hardware required to allow VSFL for ancillary service. Recall that in the experiments surveyed here, this communication was uni-directional, so the overhead is not large.

A more modest goal is to limit communication to local frequency measurements. With this limited information and extremely simple control algorithms, grid inertia could be created synthetically. Because these systems are tunable, the service to the grid can go far beyond system inertia [29]. Many residential loads could be equipped with this modest intelligence.

4.3 Engineering and economics

Section 1.3 contains a summary of the goals for VSFL: Deliver inexpensive and accurate tracking to the BA, and reliable service and incentives to the consumer. For residential consumers, these goals can be obtained using contracts between consumers and the BA or an aggregator.

The formation of contracts involves many factors: The risk to both parties on short and long time-scales. Short-term risk has been the focus of this article. Long term risk for the BA includes the uncertainty of future needs. New technology may make the cost of ancillary service very low, so that the BA does not want to be forced into a rigid contract. This risk can be reduced through two-part contracts that provide a payment for engagement by the consumer, and then regular payments based on service to the grid.

References

[1] FPL on call saving program. http://tinyurl.com/k3ldwe3, 2013.

[2] Commercial buildings energy consumption survey (CBECS): Overview of commercial buildings. http://tinyurl.com/EIACBECS03, December 2003.

[3] D. Bhatnagar, A. Currier, J. Hernandez, O. Ma, and B. Kirby. Market and policy barriers to energy storage deployment. Technical Report SAND20137606, Sandia National Laboratories., Albuquerque, New Mexico, September 2013.

[4] B. Biegel, P. Andersen, J. Stoustrup, M. B. Madsen, L. H. Hansen, and L. H. Rasmussen. Aggregation and control of flexible consumers - a real life demonstration. In *Proceedings of the 19th IFAC World Congress*, Cape Town, South Africa, Aug. 2014. IFAC.

[5] C. S. B. U. Design and Engineering Services. Pool pump demand response potential: Demand and run-time monitored data. DR 07.01 Report. Prepared by: Design & Engineering Services, Customer Service Business Unit, Southern California Edison. . Technical report, Emerging Technologies Coordinating Council, June 2008.

[6] J. Cardell and L. Anderson. The influence of demand resource response time in balancing wind and load. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 2364–2370, Jan 2013.

[7] D. P. Chassin. Load control analysis for intermittent generation mitigation. In *46th Hawaii International Conference on System Sciences (HICSS)*, pages 2305–2311, Jan 2013.

[8] Y. Chen, A. Bušić, and S. Meyn. Individual risk in mean-field control models for decentralized control, with application to automated demand response. To appear, 53rd IEEE Conf. on Decision and Control (Invited), 2014.

[9] R. Couillet, S. Perlaza, H. Tembine, and M. Debbah. A mean field game analysis of electric vehicles in the smart grid. In *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, pages 79 – 84, march 2012.

[10] G. Goddard, J. Klose, and S. Backhaus. Model development and identification for fast demand response in commercial HVAC systems. *IEEE Transactions on Smart Grid*, 5(4):2084–2092, July 2014.

[11] E. R. Hamilton, J. Undrill, P. S. Hamer, and S. Manson. Considerations for generation in an islanded operation. *IEEE Petroleum and Chemical Industry Technical Conference*, pages 1–10, Oct. 2009.

[12] H. Hao, A. Kowli, Y. Lin, P. Barooah, and S. Meyn. Ancillary service to the grid through control of fans in commercial building HVAC systems. *IEEE Transactions on Smart Grid*, 2013. under review.

[13] H. Hao, T. Middelkoop, P. Barooah, and S. Meyn. How demand response from commercial buildings will provide the regulation needs of the grid. In *50th Allerton Conference on Communication, Control, and Computing*, pages 1908–1913, 2012.

[14] B. Kirby and E. Hirst. Generator response to intra-hour load fluctuations. *IEEE Trans. on Power Systems*, 13(4):1373–1378, Nov 1998.

[15] E. M. Krieger. *Effects of variability and rate on battery charge storage and lifespan*. PhD thesis, Princeton University, http://tinyurl.com/KriegerBatteryThesis13, 2013.

[16] Y. Lin, P. Barooah, and S. Meyn. Low-frequency power-grid ancillary services from commercial building HVAC systems. In *IEEE Conference on Smart Grid Communications (SmartGridComm)*, pages 169–174, Oct 2013.

[17] Y. Lin, P. Barooah, S. Meyn, and T. Middelkoop. Experimental evaluation of frequency regulation from commercial building HVAC system (submitted). *IEEE Trans. on Smart Grid*, 2014. http://tinyurl.com/HVACtoGRID2014-pdf.

[18] Y. Lin, S. P. Meyn, and P. Barooah. Low-frequency power-grid ancillary services from commercial building HVAC systems. In *Proceedings of IEEE SmartGridComm*, October 21-24 2013.

[19] Z. Ma, D. Callaway, and I. Hiskens. Decentralized charging control for large populations of plug-in electric vehicles: Application of the Nash certainty equivalence principle. In *Control Applications (CCA), 2010 IEEE International Conference on*, pages 191 –195, Sept. 2010.

[20] J. MacDonald, P. Cappers, D. S. Callaway, and S. Kiliccote. Demand response providing ancillary services a comparison of opportunities and challenges in the U.S. wholesale markets. In *Grid-Interop*, Irving, TX, December 2012 2012.

[21] J. Mathieu, S. Koch, and D. Callaway. State estimation and control of electric loads to manage real-time energy imbalance. *IEEE Trans. on Power Systems*, 28(1):430–440, 2013.

[22] J. L. Mathieu. *Modeling, Analysis, and Control of Demand Response Resources*. PhD thesis, University of California at Berkeley, 2012.

[23] S. Meyn, P. Barooah, A. Bušić, Y. Chen, and J. Ehren. Ancillary service to the grid using intelligent deferrable loads. *ArXiv e-prints: arXiv:1402.4600*, 2014. *IEEE Transactions on Auto. Control*, accepted for publication.

[24] Midcontinent Independent System Operator. MISO business practice manual 2, energy and operating reserve market. http://tinyurl.com/MISOmanual, 2013. Last accessed: Nov 2013.

[25] PJM Website. Market-based regulation. http://tinyurl.com/mxgoagz. last accessed: June 15, 2015.

[26] D. Todd, B. Helms, M. Caufield, M. Starke, B. Kirby, and J. Kueck. Providing reliability services through demand response: A preliminary evaluation of the demand response capabilities of Alcoa Inc. Technical report, LBNL and Alcoa, January 2009. Report no. ORNL/TM-2008/233.

[27] D.-C. Tomozei and J.-Y. L. Boudec. Satisfiability of elastic demand in the smart grid. *CoRR*, abs/1011.5606, 2010.

[28] U.S. Energy Information Administration (EIA). Electricity explained - use of electricity, 2010.

[29] C. Zhao, U. Topcu, N. Li, and S. H. Low. Power System Dynamics as Primal-Dual Algorithm for Optimal Load Control, 2013.