INVITED PAPER

# Achieving Controllability of Electric Loads

*This paper discusses actively involving highly distributed loads in power system control actions; an overview of system control objectives is provided.*

By Duncan S. Callaway, *Member IEEE*, and Ian A. Hiskens, *Fellow IEEE*

**ABSTRACT** | This paper discusses conceptual frameworks for actively involving highly distributed loads in power system control actions. The context for load control is established by providing an overview of system control objectives, including economic dispatch, automatic generation control, and spinning reserve. The paper then reviews existing initiatives that seek to develop load control programs for the provision of power system services. We then discuss some of the challenges to achieving a load control scheme that balances device-level objectives with power system-level objectives. One of the central premises of the paper is that, in order to achieve full responsiveness, direct load control (as opposed to price response) is required to enable fast time scale, predictable control opportunities, especially for the provision of ancillary services such as regulation and contingency reserves. Centralized, hierarchical, and distributed control architectures are discussed along with benefits and disadvantages, especially in relation to integration with the legacy power system control architecture. Implications for the supporting communications infrastructure are also considered. Fully responsive load control is illustrated in the context of thermostatically controlled loads and plug-in electric vehicles.

**KEYWORDS** | Ancillary services; load control; power system control; power system operation

## I. INTRODUCTION

The purpose of this paper is to explore the conceptual requirements and opportunities to develop load control schemes that are competitive with conventional generation-based approaches to providing power system control services. In principle, practically any measure that can be taken by generating units (i.e., the "supply side") to ensure that electricity generation and load are equal has an equivalent countermeasure that can be taken by loads (the "demand side"). The primary characteristic of load control that distinguishes it from conventional generation-based approaches is that it must deliver a reliable resource to the power system while simultaneously maintaining a level of service commensurate with customer expectations. These two objectives are often in competition, and one of the greatest technical challenges to the competitiveness of engaging loads in power system services is to develop approaches that balance these objectives [1].

In order to balance systemic and local control objectives, we believe load control schemes must meet the dual goals of being both fully responsive and nondisruptive. In this context, we define "fully responsive" as enabling high-resolution system-level control across multiple time scales. This is desirable for load control competitiveness because conventional generation-based approaches are themselves fully responsive. Furthermore, we define "nondisruptive" control as having an imperceptible effect on end-use performance [such as building temperature, lighting levels, pump speeds, and electric vehicle state of charge (SoC)]. Nondisruptiveness underpins the reliability and cost of providing load control: strategies that are disruptive run the risk of frustrating customers to the point where they withdraw from the program or demand higher payments to participate. Though we will discuss price response in this paper, we will focus on direct load control and how organizations such as distribution utilities or third-party companies might aggregate loads to achieve a desired response, possibly for integration into electricity markets.

Manuscript received December 14, 2009; revised June 10, 2010; accepted September 7, 2010. Date of publication November 22, 2010; date of current version December 17, 2010. Research supported by the Michigan Public Service Commission through Grant PSC-08-20, and the National Science Foundation through EFRI-RESIN Grant 0835995.

**D. S. Callaway** is with the Energy and Resources Group and the Department of Mechanical Engineering, University of California, Berkeley, CA 94720-3050 USA (e-mail: dcal@berkeley.edu).

**I. A. Hiskens** is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: hiskens@umich.edu).

Digital Object Identifier: 10.1109/JPROC.2010.2081652

Despite the challenges identified above, several key advantages follow from using loads for system services.

1) Although individual loads may become unavailable at any moment, the variability of the total contribution of a very large number of small loads is likely to be less than that of a small number of large generators (for which the failure of one can have substantial impact on the ability to provide the desired service) [2].
2) Loads can often respond to operator requests instantaneously, whereas generators require some time to make output changes of any significance [2].
3) Because loads are distributed throughout the grid, they provide the opportunity to devise spatially precise responses to contingencies.
4) In some situations, using loads to provide system services could reduce overall grid emissions (for example, if relatively inefficient but fast ramping generation is no longer required to balance grid variability) [1].
5) The level of spatial and temporal flexibility that loads could provide to the power system might be used to support the growing penetration of intermittent renewable electricity generators [3].
6) Loads are already embedded in the power system, and versatile communications platforms—ranging from broadband internet connections to advanced metering infrastructure (AMI)—are becoming widely available. It may soon be the case that the only technical impediment to reliable utilization of loads for system services is the development of the necessary load models and control strategies [4].

While this paper provides a broad discussion of load control topics, where necessary, we will focus on control schemes that could be used to access small loads in residential and commercial buildings. This is because 1) these loads are ubiquitous and, if utilized in large numbers, could provide substantial, reliable system services with limited end-use disruption, and 2) with the aforementioned deployment of communications platforms, control access to these loads could be very inexpensive. We note that utilizing AMI presents a number of challenges, most notably regarding the extent to which conventional telemetry equipment can be replaced, and data/infrastructure ownership issues. We will discuss these challenges and their implications on the design of control strategies in this paper.

The paper proceeds as follows. Section II reviews the current supply-side grid operating paradigm, and provides an overview of existing efforts to integrate the demand side. A framework for achieving fully responsive, nondisruptive load control is developed in Section III. Section III-B motivates our focus on direct load control rather than price responsiveness. Control architectures are discussed in Section IV, and the related communications requirements are explored in Section V. Examples of fully responsive load control schemes are introduced in Section VI, and conclusions are provided in Section VII.

## II. POWER SYSTEM OPERATION AND CONTROL

In this section, we will discuss load control functions by first describing the major components of the conventional supply-side power system control paradigm. Then, for each major function, we will discuss efforts to integrate loads into these paradigms. We will place an emphasis on frameworks for control that have been developed in the literature as well as some of the state-of-the-art demonstrations of modern load control.

We are less concerned, in this paper, with applications of load control during emergency operation. The general concepts can, however, be extended naturally to allow load control to assist in alleviating voltage collapse [5]. The use of load control for damping oscillatory modes [6] is also beyond the scope of this paper.

### A. Economic Dispatch and Unit Commitment

*1) Supply-Side Paradigm:* Electricity demand in a power system varies throughout the day, following patterns that depend on, among other things, regional characteristics, temperature, time of day, day of week, and season of the year. Decisions to change generator output to accommodate variation on hourly time scales are usually made by processes of *unit commitment* and *economic dispatch* [7]. Unit commitment establishes generator operating schedules in advance of the operating time and takes into account generator ramping capabilities and startup and shutdown costs. Solutions are obtained via a multiperiod optimization process such as dynamic programming, Lagrange relaxation, or mixed integer programming. Unit commitment determines when to bring generators online and offline, and so is typically run one day in advance. Economic dispatch is the process of choosing the output levels for generators that are already online, with the objective of minimizing the total cost of meeting demand. Economic dispatch tends to be quite fast, and can be run within minutes of the operating time. Both processes require demand forecasts.

*2) Demand-Side State of the Art:* In the supply-side paradigm above, as electricity demand increases and increasingly inefficient and expensive generation is brought online, generation costs can become so high that supply-side costs exceed the retail price by an order of magnitude or more. Therefore, for several decades, many utilities have maintained the infrastructure to curtail electricity loads (especially air conditioners and water

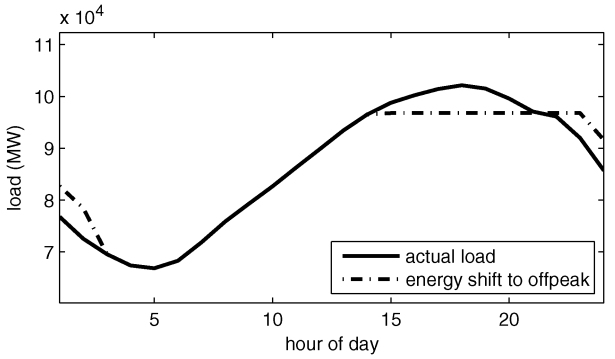


**Fig. 1.** *A hypothetical redistribution of load from peak to off-peak hours. (Original data taken from the Midwest Independent System Operator website, www.midwestiso.org.)*

heaters) to reduce load rather than dispatch additional generation during periods of very high demand. Fig. 1 shows a desirable redistribution of load away from the peak. Load pattern changes such as this are intended to reduce supply side operating costs (by reducing the need to build and operate high marginal cost peaking generation) and improve system reliability (by maintaining an acceptable operating reserve).

The central challenges associated with directly controlling electrical loads to contain generation costs are that:

1) the total power and energy available for control is limited by the obvious need to serve the primary end-use function of the load;
2) there could be a postcontrol "pickup" (or "recovery peak") in load that results from the continuous operation of previously controlled loads as they recover their desired operating state (e.g., temperature setpoint). In some circumstances, the recovery peak can cause total load to exceed that which would have occurred in the absence of control.

Over the years, numerous papers have been written to address these challenges. In all cases that we are aware of, the primary control mechanism is to curtail load operation during high demand hours and manage the subsequent recovery peak (such as in Fig. 1). It is common to use simple energy "payback" models that do not model the specifics of load end-use function but instead model the recovery peak simply as a redistribution of the system-wide curtailed energy during the control interval. Many papers focus on developing strategies to produce a predefined load trajectory [8]–[12]. Others treat the load trajectory as endogenous to the model by integrating payback models into system operator economic dispatch and unit commitment processes [13]–[18]. Although some papers do attempt to incorporate the end-use function of the load into the control decision-making process [19], researchers typically use off-time as a proxy for end-use function. However, recent papers have started to integrate detailed thermal load models into load curtailment decisions [20]–[22].

In the papers cited above (with the exception of [21] and [22]), as well as in many implementations of load management schemes, control is achieved by engaging relays that interrupt power to loads. Those relays are usually activated by the utility or load serving entity via a radio signal, telephone, or modulated carrier signal sent directly over the power lines [23].

Load control for peak shaving in modern implementations is undergoing significant transformations, with increasing competition coming from "third party" aggregators who serve as intermediaries between the loads and a system operator [24]. In this case, an aggregator (which could be a utility) submits curtailment bids to the system operator; if the bid is accepted the aggregator curtails a group of loads under its control to achieve the reduction [25]. In some cases, aggregators are using thermostat temperature setpoint as the control input.

Communications with loads are becoming increasingly sophisticated. One notable development has come out of efforts at Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA. The "OpenADR" project at LBNL has developed automated demand response procedures for peak load reduction, and these are now being adopted by utilities and their customers [26]. The core concept of OpenADR is a communications platform that uses open internet protocols which are interoperable with different building and industrial control systems. Via this platform, utilities can indicate to their customers' energy management systems (EMSs) when the grid is operating at or nearing capacity. Each EMS is then capable of delivering an automated but customized and overrideable response to the utility request. The major innovation of this approach is its ability to interface with a variety of customer control systems.

Many existing peak demand management programs that utilize direct load control are disruptive and can have significant impacts on the end use. In the case of air conditioning, in most regions of the United States, load is highest on the warmest days, meaning that air conditioners are curtailed when their services are most in demand. Though the research cited above explores the use of feedback control, current large-scale implementations are open loop and relatively unsophisticated with respect to minimizing impact on the end-use function. This is at least in part due to the historically high cost of reliable sensing equipment; AMI and other developments that rely upon advanced communications platforms may change this situation.

Plug-in electric vehicles (PEVs) are appealing as controllable loads because they could be curtailed for significant periods of time (e.g., several hours) without impact on end-use function. In fact, they even offer the possibility of returning power to the grid. Provided that a

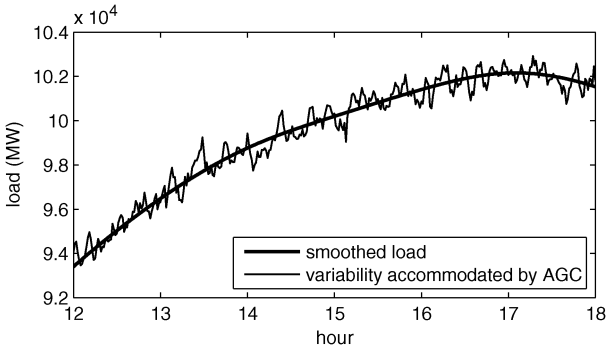


**Fig. 2.** ***An example of the rapid time-scale variability that loads or fast-responding generators might mitigate via automatic generation control.***

vehicle's battery SoC is sufficient at the time it is needed for on-road use, the vehicle owner has little concern for the details of when and how quickly it is charged. (An exception to this is related to the impact that charging rates may have on battery state of health and its lifetime.) PEVs could be managed not only during peak hours (when their contribution to load may not be very large) but also for night time "valley filling" load control strategies that distribute PEV charging to minimize total energy costs. We will discuss this in more detail later in the paper.

### B. Frequency Restoration Mechanisms

*1) Supply-Side Paradigm:* Although considerable effort goes into predicting electricity demand so that generators can be dispatched as efficiently as possible, it is impossible to predict demand with complete accuracy. Second-to-second and minute-to-minute fluctuations, such as those illustrated in Fig. 2, are especially difficult to foresee and result in a difference between nominal demand and the generation scheduled to meet that demand (this difference is sometimes referred to as energy imbalance). As we explain below, these energy imbalances lead to changes in system frequency that are usually met by changing the output of flexible generators. A combination of three supply-side mechanisms typically operate in unison to achieve the desired frequency control.

First, an unanticipated change in load or generation is initially compensated for by the addition or extraction of kinetic energy from the rotating inertia of all synchronous generators; this results in a change in system frequency [27]. Second, many generators are equipped with frequency responsive governors that produce an output change proportional to the frequency deviation (the constant of proportionality is known as a speed-droop characteristic). If system frequency deviates sufficiently far from its setpoint (e.g., 35 mHz or more [28]), droop is activated to prevent further growth of the deviation. This control strategy is inherently decentralized and robust to small disturbances. Furthermore, it is initiated almost instantaneously, although a governed generator may require some time to achieve the output level dictated by its droop characteristic.

These first two mechanisms are fully decentralized and therefore not well suited to restoring system frequency to its setpoint. Instead, a third mechanism called automatic generation control (AGC) serves this purpose.[1] AGC decision-making occurs at the level of "balancing authorities" (BAs), which are relatively large regions that might contain hundreds of thousands or millions of customers. When BAs are interconnected, unanticipated changes in load or generation can result in deviations in scheduled interauthority tie-line flows as well as frequency deviations. Because both deviations are undesirable, AGC calculations are usually based on a weighted sum of system frequency and unscheduled power flows. The resulting signal is called area control error (ACE) [7]. To minimize ACE, AGC issues raise or lower signals based partly on each generator's ability to provide the desired response in a reasonable amount of time, and partly on real-time economic dispatch [7]. These signals are typically pulses of varying length (and proportional to the requested output change) that are conveyed on a dedicated communications infrastructure which also telemeters the state of all generators in the BA. Although the signals may be updated based on system ACE and issued as frequently as once every two seconds, economic dispatch targets will not be updated that frequently due to the required computing time.

AGC generally relies solely on instantaneous generator availability and ACE signals (although a model predictive control approach to AGC has been developed in [29]). Engaging loads in the AGC process may, however, require control strategies that forecast how loads will respond to control signals in the future.

*2) Demand-Side State of the Art:* Energy imbalance is driven by forecast errors which are, for an unbiased forecast, roughly zero mean. For this reason, electricity loads with some form of energy storage (either thermal in the case of thermostatically controlled loads, or electrical in the case of PEVs) are excellent candidates for imbalance control. This is because their average power consumption under imbalance control can be made to equal their nominal consumption in the absence of control.

Many loads are "energy constrained" in the sense that they will cease to provide their primary end-use function if they do not receive a sufficient amount of energy over some time interval (i.e., if their average power consumption is too low or high). If the mean AGC response provided by a load approaches zero over a relatively short averaging time frame, or if its storage capacity is large, the

[1]AGC is sometimes referred to as load frequency control or regulation.

load's energy constraint is likely to be achievable. On the other hand, if a large response relative to the load's energy capacity is desired, some form of dynamic optimization may be required to minimize the impact on the end-use function of the load. In this case, in contrast to the supply-side paradigm, a multiperiod optimization process would be required.

There are some early stage efforts to manage energy imbalance with frequency-responsive load by providing the equivalent of generator droop [30]–[32]. As with generator droop, the approach is completely decentralized, but in the case of loads it may be challenging for system operators to predict or verify the system-level response that will be produced by thousands or millions of unknown devices. An alternative is to integrate loads into a centralized AGC-type scheme, because it will provide the system operator with awareness and control of the response and facilitate restoration of system frequency to its nominal value.

We are aware of only two efforts to study the provision of AGC with aggregated loads by some type of centralized load control. One of these efforts is by one of this paper's authors [33], and addresses the ability to provide responses on 1-min intervals with thermostatically controlled loads. A second effort, at LBNL, is exploring the potential to use dimmable lighting for regulation [34].

Regulatory structures exist for large-scale provision of frequency control with loads, but as yet there has been relatively little actual implementation. The U.S. Federal Energy Regulatory Commission issued Order 719 in October 2008. This order requires independent system operators (ISOs) and regional transmission organizations (RTOs) to allow demand resources to participate in ancillary services markets, including AGC. This has led to a number of studies and task forces at the United States ISOs and RTOs. However, most programs that could support AGC are very recent and as yet there is relatively little actual participation of loads [35], [36]. There are a number of reasons for such limited participation, including unfamiliarity with a new service, minimum size, and the requirement that the loads themselves (rather than an aggregator) be capable of receiving and responding to AGC commands [36]–[38]. In the case of the Electric Reliability Council of Texas (ERCOT), power flow at the point of metering must actually be net *generation* in order to satisfy telemetry requirements [39]. Because most ISOs and RTOs are establishing telemetry requirements equivalent to those of generation facilities, it is likely that only large loads will ever be capable of making the investment required to participate under the current regulatory environment.

### C. Contingency Reserves

*1) Supply-Side Paradigm:* When a sudden, large loss of power supply occurs on the grid (for example, a generator or transmission line trips offline), a large frequency excursion occurs. That causes frequency-responsive generators (referred to as *spinning reserve*) to automatically begin increasing their output to reduce the supply imbalance. Following such an event, it is common for AGC to be disabled until the system operator is able to restore grid frequency (or ACE) to its setpoint by manually issuing raise-lower signals to reserve capacity via the system telemetry infrastructure. This might take 5–10 min as spinning reserve generators cannot instantaneously increase their output. In order to have sufficient capacity to quickly accommodate a contingency, spinning reserve generators must be grid connected and operating in a part-loaded state. Part-load operation is usually inefficient, so spinning reserve increases operating cost and emissions.

As with AGC, spinning reserve generation need not be dispatched via a multiperiod optimization process. Although these generators may be limited in how long they are capable of providing reserve power, the duration of this limitation is typically not binding as system operators can usually bring supplemental reserves online in less than an hour.

*2) Demand-Side State of the Art:* Electricity loads are well suited to providing reserves because they can respond very quickly (in many cases the ramp rate is constrained only by the speed of the communications network). For some time, system operators have used nonselective load shedding (i.e., disconnecting entire regions from the grid) as a measure of last resort to avoid system collapse. Selective load shedding (i.e., disconnecting customers or specific customer loads based on prearranged agreements), on the other hand, has much more potential from the perspective of customer acceptance because noncritical loads can be targeted for shedding. As in the case of using loads to manage energy imbalance, those with significant energy storage capacity (thermal or electrical) are especially well suited for providing spinning reserve. This is because the time required to restore the system, and allow loads to return to normal, is often short enough that the end-use function may not suffer [40]. A number of recent publications and white papers have explored the potential of using responsive loads for spinning reserve [2], [17], [18], [32]. Furthermore, demonstration projects with relays on residential loads—originally installed for peak load shaving—are showing promising results [40].

Several electricity markets (including ERCOT, ISO New England, and PJM in the United States, and systems in the United Kingdom, Norway, Finland, and Australia) have instituted programs to use loads as reserves [37], [41], [42]. These programs typically focus on large industrial loads. Furthermore, in many cases, the control strategy is relatively crude. For example, in the case of ERCOT, contingency reserves can be activated in one of two ways. First, for smaller events, "voice dispatch" is practised (i.e., the system operator picks up a telephone). Second, if the event is large enough and frequency drops by roughly 0.5%

of nominal, loads with underfrequency relays will automatically trip. Because the system response to a large quantity of load tripping under these circumstances is not well understood, ERCOT limits the total capacity of load that can participate in spinning reserve activities to half of their total reserve capacity. The program is fully subscribed [39].

Several third party aggregators are entering the contingency reserves market [25], [43]. In contrast to the ERCOT approach of dispatch by voice, these organizations are using advanced communications equipment to activate loads.

## III. ACHIEVING FULLY RESPONSIVE LOAD CONTROL

As described previously, our objective in this paper is to explore frameworks that make loads competitive with the supply side for providing system-level services. In this regard, we have defined the goal that load control be fully responsive (enabling high-resolution system-level control across multiple time scales) and nondisruptive (control action has an imperceptible effect on end-use performance).

In general, the existing demand-side strategies discussed in the previous section fall short of this goal. For example, with the exception of [21] and [22], the body of load shifting work above controls local power but does not monitor end-use performance. This makes the goal of nondisruptive load control harder to achieve than if local conditions were monitored and/or used as the control input. In the case of frequency restoration mechanisms, decentralized schemes, though potentially useful, do not facilitate system-level objectives. The efforts to provide spinning reserve with loads may be closest to achieving the goal of nondisruptive control, since curtailment events can be short enough to avoid disruption. However, this addresses only one time scale of control.

In this section, we explore some of the key conceptual issues that influence the achievability of fully responsive nondisruptive load control.

### A. Dual Objectives of Load Control

With advanced metering infrastructure and other emerging grid "cyber-infrastructure" developments, it is becoming increasingly feasible to control loads to provide the system services discussed above. At least two challenges must be addressed for this to succeed. First, strategies for incorporating load control into power system operations must be consistent with the legacy system and responsive to system-level requirements. Second, load control schemes must achieve end-user acceptance. Without acceptance, loads cannot be recruited for control. Worse yet, if acceptance erodes over time, customers who were previously recruited may withdraw from a load control program, possibly when their capacity is needed most, such as during a contingency event. This issue, known as "response fatigue" [44], is central to load control program design and its long-term value.

Therefore, as we stated in the introduction, one of the central challenges facing a fully responsive load control program is delivery of a reliable resource to the power system while maintaining a level of service for end users. In other words, effective approaches to load control must balance the dual objectives of two levels of control. These levels of control could frequently be in competition. This is almost always the case in peak-load shifting programs that rely on air conditioning loads as the primary source of controllable capacity (where system-level load needs to be shifted most on days when air conditioning is most in demand).

These requirements suggest the need to take end-use function into consideration as load control decisions are made. An intuitive strategy would be to control loads whose end-use function will be least affected by the control action. For example, one might choose to selectively control thermal loads that are closest to their desired temperature setpoint, since they can absorb a greater service disruption than others. This can be done either by 1) dictating the power consumption of each device but basing control decisions on feedback from loads' end-use function status, or 2) by directly controlling the end-use function (e.g., temperature setpoint) with the expectation that power consumption will change according to some known model.

As with the services described in the previous section, there are challenges associated with coordinating thousands or even millions of loads in a way that minimizes end-use impact, or guarantees a certain level of end-use function [45], [46]. These challenges stem from, among other things, information and communications bandwidth requirements, model fidelity, and controller design. These issues are addressed later in the paper.

### B. Price Response Versus Direct Control

Methods for engaging loads in power system services can be distinguished by whether they issue a signal that reflects instantaneous electricity generation costs. Price-based signals are effective inputs for integrating loads in longer time scale economic dispatch functions, and are comprehensively reviewed in [47]–[49].

We will not, however, consider price response as a mechanism for achieving fully responsive nondisruptive control for several reasons. First, electricity markets do not presently clear on time scales faster than 5 min. Consequently, price signals are not used for fast services such as regulation and spinning reserve on the supply side. (We note that the 5-min threshold between price-based and nonprice-based load response dates back at least as far as the seminal work of Schweppe *et al.* in 1980 [47]). Second, having direct control over loads increases the system operator's ability to predict the loads' responses (though price response forecasts certainly are possible),

and provides third-party aggregators certainty over how much capacity they can bid into ancillary service markets [43]. Finally, customers, especially small ones, may be disinterested in (or incapable of) identifying their own demand curve (i.e., instantaneous quantity responsiveness as a function of real-time price) if their objective is to receive a service that is a function of energy use over time (e.g., thermal comfort) rather than instantaneous consumption.

Instead, in this paper, we will focus on control strategies that employ a nonprice-based signal (perhaps a "functional" quantity like temperature setpoint or a signal that directly dictates power consumption). To ensure customer acceptance, this type of direct control will need to be capable of guaranteeing some level of end-use function, and we will explore mechanisms to do so in cases where communications bandwidth is constrained. Such strategies could make use of financial incentives to encourage customers to consider greater flexibility in their quality of service (as in so-called incentive-based programs [50] or with interruptible service contracts [51], [52]).

It is worth noting that one of the main criticisms of nonprice responsive load control is that customers often receive payment for load reduction from some baseline, yet the baseline is impossible to directly measure [53]. This is an important issue for load curtailment over longer (e.g., multihour) time scales when the baseline is a time-varying quantity that needs to be modeled, and at the very least there is a need to develop reliable baseline modeling methods [54]. On the shorter time scales for AGC and spinning reserve, however, the quantity of interest to the system operator is the change in consumption from one instant to the next. This can be measured, either by real-time telemetering equipment (which would likely reside at the substation level and measure power changes in aggregate) or by a local gateway, which could send data to a central coordinator for verification purposes in real time or after the fact. We note that in its current form, AMI does not record data quickly enough to serve this purpose.

### C. Motivating Examples: Uncoordinated Load Control

In assessing the role and value of fully responsive load control, it is helpful to consider alternatives that offer more autonomous control of loads. The examples presented in this section use time-based and price-based control strategies and show that the lack of coordination inherent in autonomous control may lead to unexpected and undesirable collective behavior.

In both examples, we consider PEV populations with required charge for each vehicle uniformly distributed over the range 10–20 kWh, and vehicle charge rates uniformly distributed across a range of 3–6 kW.

The first example considers time-based control of 4 million PEV loads. Fig. 3 shows the underlying (non-PEV) demand over a daily cycle, together with the total demand

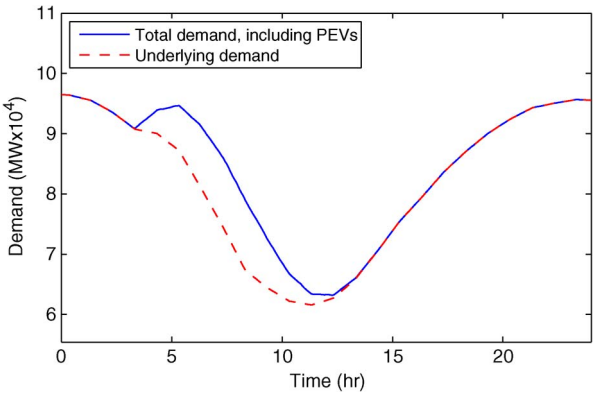


**Fig. 3.** *Total load demand due to time-based control of 4 million PEV loads. (Original data taken from the Midwest Independent System Operator website, www.midwestiso.org.)*

when PEV loads are included. The PEV charging loads are set to switch on between approximately 3.5 and 8.5 h after the original peak, with turn-on times uniformly distributed over that range.

It can be seen that this control strategy partially fills the evening load valley, but has the undesirable outcome of creating a second load peak around 6 h after the original peak. Conceptually, customers could be assigned switch-on times that were more widely distributed. Such a strategy would help moderate the PEV-induced peak, but would be somewhat inconsistent with the underlying desire for autonomy. Also, keep in mind that switch-on times are already spread over a 5-h period.

Price-based control strategies can take numerous different forms. This example uses a hysteretic form of control, whereby the charger switches on when price falls below a lower threshold and switches off when the price rises above an upper threshold. A range of PEV population sizes, up to 4 million PEVs is considered. The PEVs face the hypothetical supply curve shown in Fig. 4. The switch-on price for PEVs was uniformly distributed over the range \$60–70 per MWh, while the switch-off price for each PEV was uniformly distributed between \$3 and \$10 above its

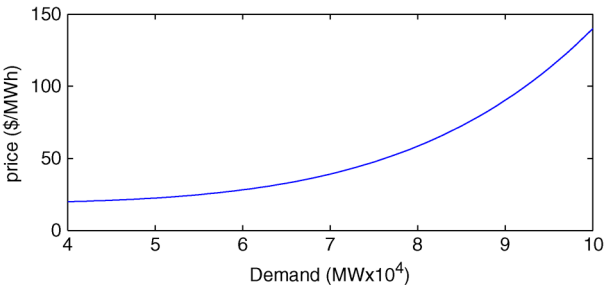


**Fig. 4.** *Hypothetical supply curve for PEV charging, with price* $= 0.0012D^5 + 18$, *where* $D$ *is total demand in MW.*

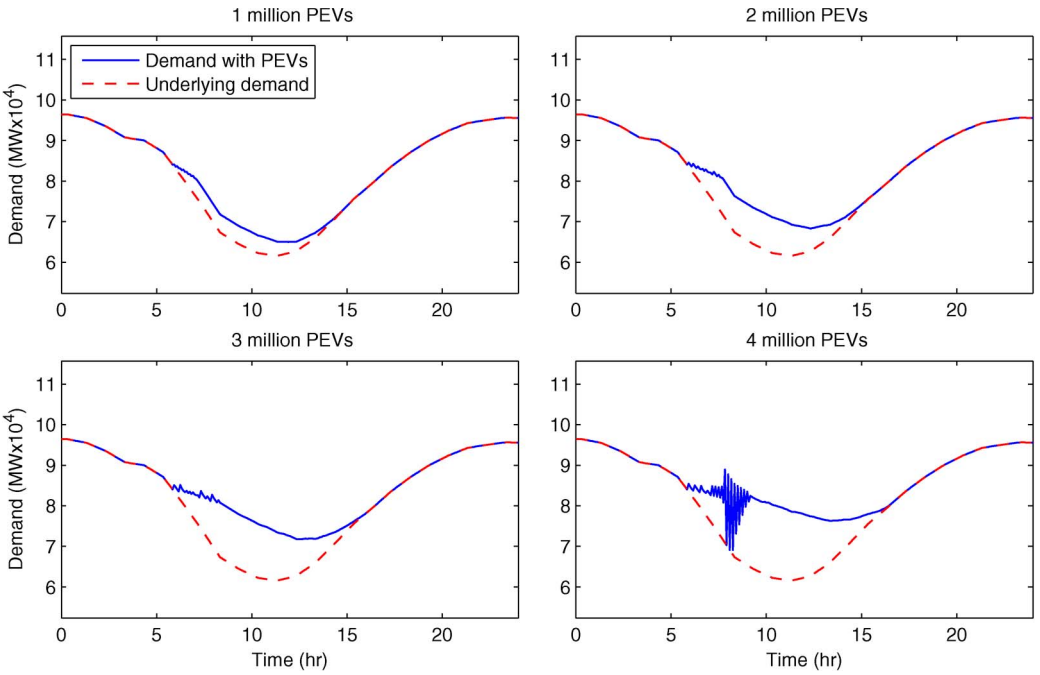


**Fig. 5.** *Total load demand due to price-based control of varying numbers of PEV loads.*

switch-on price. Resulting demand patterns for different PEV population sizes are shown in Fig. 5.

The dominant feature in Fig. 5 is the spontaneous emergence of significant oscillations around the 7-h mark for the largest population size. This oscillatory process is driven by interactions between energy price and demand. An increase in total demand due to PEV load causes an increase in energy price, which may be sufficient to curtail some of the PEV load. Interestingly, these oscillations are not present for the smaller PEV population sizes, suggesting that this form of control may be sufficient for small numbers of PEVs, but that as the number of vehicles increases more sophisticated control strategies may be needed.

Furthermore, the oscillations are not always present even for the larger population size. Outcomes are dependent upon the random selection of PEV switch-on prices, with the likelihood of spontaneous oscillations strongly influenced by the width of the switch-on price range. With a tight range, demand varies greatly for a relatively small change in price. As the price range widens, the sensitivity of demand to price reduces. This sensitivity can be thought of as an important gain in this load scheduling process. The jitter and oscillations apparent in Fig. 5 suggest a gain that is too high.

If load control is truly autonomous, the switch-on price band cannot be shaped *a priori*, but rather is determined by customer choice. Numerous factors may influence that choice, from gasoline prices to travel plans for the following day. In any case, price-based control strategies face challenges in achieving an adequate level of control while ensuring robustness to the uncertainties inherent in customer behavior.

### D. Load Control Metrics

The potential contribution of a load to a control request can be described in terms of its *availability* and *willingness* to respond to the request. The first of these metrics refers to the amount of load available for switching in or out by the control action. The measure of willingness, on the other hand, weights that load to reflect the impact of the control action on the end-use function. Consider the case of a residential customer who has a number of appliances available for control. A local customer-based load manager would monitor the status of those devices, and assess the total load available for switching, both in and out. Furthermore, each device would communicate its willingness to be switched. For example, refrigeration would be very willing to be switched out when nearing the end of its cooling phase, but quite reluctant had cooling only just begun. The load manager would consolidate all the appliance information into a load model suitable for higher level coordination by an aggregator or utility. Such a scheme is presented in [46], where the willingness measure is quantified in terms of a price function.

The willingness measure forms the basis for mapping between the controllable loads and the control services that they can most appropriately provide. Loads that were most willing to participate would be well suited to AGC regulation. Less willing loads could be called upon to provide contingency reserves.

The concepts of availability and willingness describe the immediate state of load controllability, but do not facilitate the management of temporal constraints. Control decisions might be made that leave insufficient controllability over time. There would seem to be benefits in loads providing information describing their energy requirements and delivery time frame. It would then be possible to build, and continually update, a load control schedule that maximized controllability over a finite horizon while satisfying energy delivery constraints. A challenge lies, however, in communicating those constraints to the system operator in a way that leads to control decisions that maximize the use of loads without compromising end-use function.

Centralized acquisition of the amount of information needed to achieve precise controllability of every individual load would require substantial communications and processing resources. Load control strategies that avoid such extensive cyber-infrastructure are considered in Section IV. An alternative to acquiring information from every load is to randomly sample the loads, and build a statistical assessment of load controllability [55].

### E. Choice of Input Signal

In current operating paradigms, control of power system elements involves requesting changes in quantities that are of direct relevance to the operating state, e.g., real or reactive power, or terminal voltage. Load control could follow this route by choosing real power as a control input signal. Because most loads have relatively little inertia (unlike rotating generators), power responses could be fast and precise, even if thousands or millions of loads are under control. However, to accomplish nondisruptive control, such a strategy would require knowledge of the relationship between power consumption and end-use function. Because each load is subject to slightly different conditions and has slightly different characteristics, it could be necessary to model or measure the state of every device in the population under control. Without significant device-level monitoring, this approach risks poor end-use performance in some loads, and possible response fatigue.

Alternatively, a functional quantity, such as temperature, lighting intensity, or PEV battery SoC could be chosen as the input signal. Then, assuming that the control mechanism at the end user is operating properly, it becomes less important to monitor the state of the loads. The challenge shifts to estimating how the load power will respond to the input signal. Making these predictions at the level of individual loads would be computationally challenging without detailed knowledge of each load's characteristics. It could instead be possible to predict the aggregate response with a reduced form model, if all loads receive the same input signal [33]. This type of approach appears to present less risk of response fatigue, but greater uncertainty over the aggregate power response, though system diversity reduces some of this uncertainty. Section V-B provides a discussion of issues concerning aggregate power feedback.

## IV. CONTROL ARCHITECTURES

### A. Centralized Load Control

A centralized load control strategy would require that the power system operator issue command directly to individual loads. Though this is the present practice for generation control, it would obviously be significantly more challenging to achieve for thousands or millions of loads. Furthermore, as discussed above, feedback—either on the end-use function or power demand—would likely be required to achieve reasonable performance. If response fatigue is an issue, using a functional quantity as the input signal could be a better option, because using power as the input signal would require monitoring the load functional states. This would be impractical for a large system. Issuing control signals (especially if they are functional quantities rather than power) for millions of devices would not, however, fit well into the legacy system.

### B. Hierarchical Load Control via Aggregators

Rather than have millions of devices interact with high-level power system controls as in the centralized model, load aggregators (which could include third-party firms or load serving entities) can serve as intermediaries, as shown schematically in Fig. 6. This hierarchical structure provides a framework under which load aggregators could

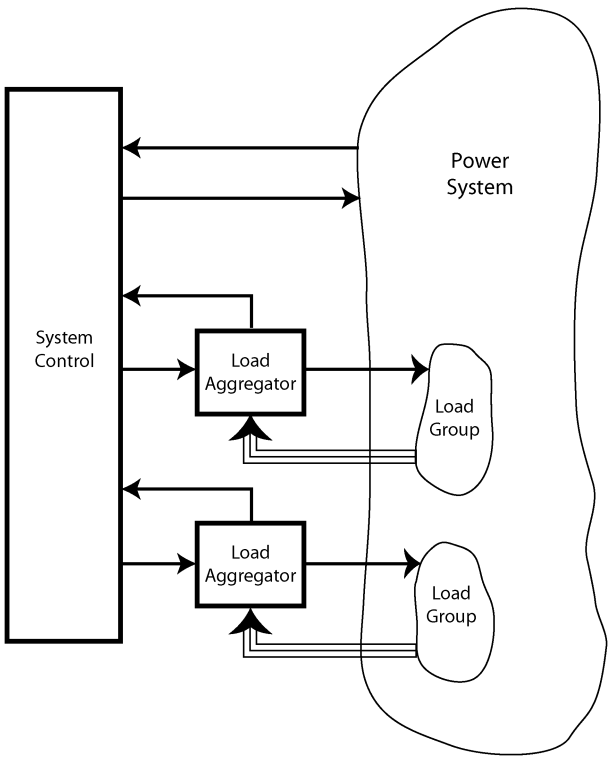


**Fig. 6.** *Schematic representation of a hierarchical load control strategy.*

engage a very large number of loads within the legacy operating paradigm.

In this model, each load aggregator has jurisdiction over a certain group of loads, and provides an interface between those loads and the higher level controls. It acquires information from participating loads which describes their availability and willingness to respond to control actions, as discussed in Section III-D. The aggregator can then use the information provided by individual loads to build a model of the responsiveness for the entire group. The exact form of that model depends on the role that the group may be called upon to perform. For example, in the case of AGC, the model would describe the load increase/decrease achievable in the short term.

To seamlessly integrate into the existing system, the aggregator should appear as a "virtual power plant" [22] to the higher level controller, and accept the same commands (e.g., raise/lower) as those received by an individual generator. Aggregators should then be capable of dispatching their loads to respond to the higher level commands. This implies that the aggregator must interpret the control signals received from the higher level controller, and pass on instructions that are meaningful to the loads.

Aggregators may not have the ability to tailor instructions for individual loads. To do so would require each aggregator to maintain a detailed database that dynamically updated to accurately reflect the composition of the load group. A simpler approach would be for the aggregator to broadcast a common signal to all loads in the group, allowing the loads to interpret that signal and respond accordingly. The examples in Section VI present possibilities for such signals.

### C. Opportunities for Distributed Control

As discussed in Section II-B, the legacy control system includes frequency responsive generators that operate independently of the centralized control system. It would be relatively straightforward to enable loads to be frequency responsive as well. As a next step in the direction of decentralized control strategies, system ACE could be broadcast to loads [30]–[32].

The challenge of a completely decentralized approach is that local decisions may result in an over- or undersupply of the required response, and local controllers may work against each other to achieve their desired goal. This challenge is not new—generator droop is designed specifically to avoid conflict among decentralized controllers. However, it then becomes impossible to maintain a system-level setpoint value such as frequency with the decentralized approach alone.

This could be addressed by establishing a distributed control structure, where controllers communicate to achieve a common goal [29]. However managing the quantity of information that needs to be exchanged could be challenging, and coordination times may be excessive. There may be opportunities to collect individual load states and control decisions at a centralized location, and redistribute that information in an aggregated form [56] that is manageable and useful to loads.

## V. COMMUNICATIONS REQUIREMENTS

Each of the control architectures discussed in Section IV places its own particular requirements on the supporting communications networks. Nevertheless, all communications networks must take into account the highly distributed nature of loads. This section considers a range of issues that arise in seeking to establish coordinated control of large numbers of loads. We will focus on advanced metering infrastructure because it is being widely deployed. However, many of the issues (latency, bandwidth, data ownership) discussed have relevance beyond AMI.

### A. Infrastructure

As mentioned earlier, the most practical forms of load control tend to utilize control commands that are broadcast across all loads, rather than targeted to specific installations. Such signals could be delivered via the AMI network, though alternatives include delivery over each individual customer's broadband internet connection. The challenge of the latter option is that the communications network would be owned by another party, who is not directly involved in power system operations. This could lead to complications with maintenance, reliability, and the ability to issue high priority signals to the loads under control. Furthermore, distribution companies have an incentive to build and utilize AMI capacity because it can be included in their rate base, whereas broadband connections cannot. Finally, in practice, it may be that successful control strategies will be capable of using either communications platform. This would provide a comfortable level of redundancy in the communications infrastructure that would ensure loads could reliably engage in power system controls.

AMI takes different forms, but typically consists of a home area network that communicates with the electricity meter, a wireless local area network that collects meter information in a "cell relay," and a broadband connection for passing that meter information from the cell relay onto an AMI collection point.

### B. Bandwidth Requirements

The bandwidth required to support fully functional load control is not significant because relatively little information (e.g., a vector of temperature states and power consumption) is required to describe the state of individual loads. However, current AMI wireless local area network (LAN) communications protocols may significantly limit the rate at which data can be collected. This is because meter read requests are issued at a limited frequency, on the order of once every 5 s. If unique

requests are required for each unique meter, as is currently the case for collecting AMI electricity consumption data in some AMI networks, then it would take one cell relay well over an hour to query 1000 m (a typical LAN size). The AMI network also needs to collect electricity consumption data for billing, and it typically does this three times per day. With these frequency limitations and competing AMI network uses, the ability to collect load state information for control purposes may be limited. On the other hand, because the update rate of cell relays is on the order of seconds, if they can be configured to collect data from all meters in their LAN once per update, then the speed of information gathering would be more than adequate.

To institute closed-loop feedback control in either the centralized or hierarchical control strategies described above, power measurements must be acquired at some level and communicated to the central operator or aggregator on relatively fast time scales, perhaps on the order of seconds. If cell relays are unable to collect data from all meters in their LAN once per update, it might be possible to use aggregated data from distribution substation telemetry equipment. However, the loads in the control population may be only a fraction of the capacity connected to the substation. Therefore, it will be necessary to implement some type of filtering to extract the response of the aggregated loads from variation in the rest of the substation loads.

If a common signal is broadcast to all loads, computational requirements associated with signal construction are not likely to be significant. Given that data transfers should be secure, the greatest computational burden may be associated with encryption/decryption algorithms.

### C. Data and Infrastructure Ownership

The issue of data and AMI ownership may need to be carefully considered in the development of load control schemes. The actual load data and AMI will be owned by the distribution company. However, the entity that handles load control functions (such as a third party aggregator) need not be the same organization. There are likely to be legal and technical issues involved in giving these third parties access to the distribution company's property. In the case of hierarchical control, such as illustrated in Fig. 6, communications between the aggregators and the system operator take the form of consolidated models. Individual loads are not identifiable, so the same data ownership issues will not arise at that higher level.

### D. Latency

It is well known that time delays within control loops can result in degraded performance and even instability [57]. Time delays in the measurement process cause the controller to operate on old information. On the other hand, time delays in the actuation process result in the control action influencing the system later than intended. In both cases, closed-loop performance will usually be degraded, especially when fast response times are required and/or frequent control updates are issued—as with spinning reserve and automatic generation control. Because load control involves highly distributed resources, time delays in the communications processes are unavoidable. However, it remains to be seen if those delays significantly undermine the performance of any load control scheme.

The load control structure presented in Fig. 6 effectively decouples the process of building aggregate load models from the use of those models. As suggested above, most of the communications delay is confined to the model building process. Consequently, latency will influence behavior primarily through the higher level controller's use of models that may be out of date. This is insignificant under normal load variation conditions. However, if the controller calls for a large load change, for example, in response to a need to deploy spinning reserve, the delay in model rebuilding may result in subsequent control actions that are inaccurate and potentially destabilizing.

## VI. EXAMPLES

In this section, we will discuss two load control applications where the control system is configured to control an end-use function (building temperature and battery SoC). These approaches are hierarchical in the sense that they involve at least two control layers, namely, the local controller, which serves the end-use function, and one or more other controller(s) whose purpose is to maintain a system level (real power) objective by adjusting the end-use function. However, in principle, the approach could be used in either the centralized or aggregator architectures discussed above.

### A. Thermostatically Controlled Loads

Thermostatically controlled loads (TCLs) comprise roughly 50% of electricity consumption in the United States [58] and represent an excellent end-use class for load control due to their ability to store energy in the form of temperature gradients. They can be deferred for limited periods of time without any appreciable loss of end-use function. TCLs might be an especially good end-use class for balancing faster time-scale fluctuations from intermittent renewable electricity generators, such as wind turbines and photovoltaic devices.

For example, Callaway [33] has developed control algorithms and theoretical results suitable for frequency restoration and short time-scale economic dispatch ancillary services with the specific goal of managing wind plant variability. The paper focuses on developing a model to map changes in thermostat setpoint to changes in power demand for large aggregated populations of TCLs. The

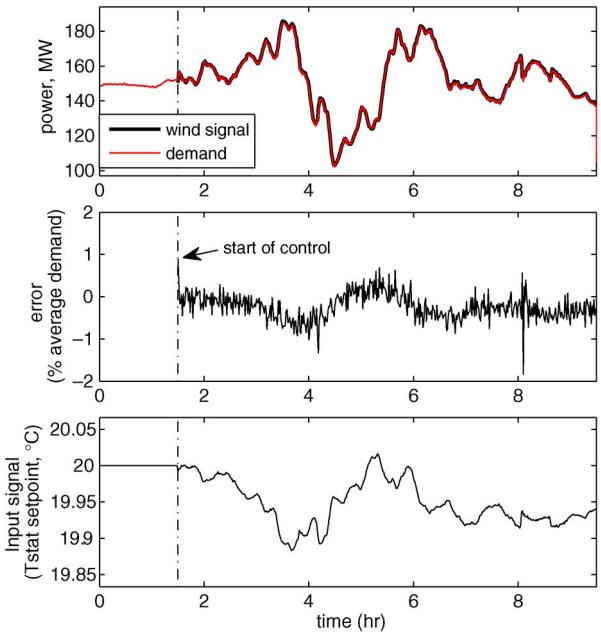


**Fig. 7.** ***Load control example for balancing variability from intermittent renewable generators, where the end-use function—in this case, thermostat setpoint—is used as the input signal. See [33] for more details.***

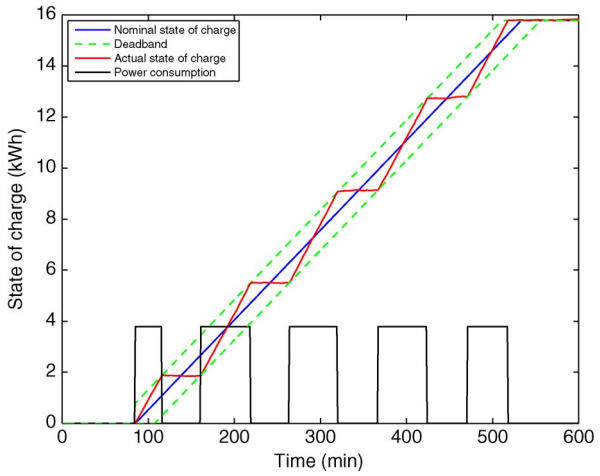


**Fig. 8.** ***Hysteresis-based PEV charging scheme.***

model can be formulated as a minimum variance controller that computes changes in thermostat setpoint required to achieve desired aggregated power responses.

Fig. 7 depicts one of the central results of the paper. The top panel of the figure shows two lines. The first is the zero-mean high-frequency component of a wind plant's output plus a direct current (dc) shift equal to the average demand of the TCL population under control. The second line is aggregate demand from the controlled population (in this case, 60 000 air conditioners), where they are subjected to shifts in their temperature setpoint as shown in the bottom panel of the figure (these shifts are dictated by the minimum variance controller). The middle panel of the figure shows the controller error, which is relatively small.

In Section III-D, load controllability was discussed in the context of availability and willingness to participate. These concepts are implicitly taken into account in the hysteretic form of control associated with thermostats. As the temperature nears either end of the deadband, a TCL becomes available for control. It becomes increasingly willing to participate in control as the temperature approaches the switching limit. However, once the TCL has switched state (encountered the deadband limit), it is temporarily no longer available for control.

Assuming relatively constant ambient temperature, the controllability of a large population of TCLs will vary little over time. However, large temperature changes affect the availability of TCLs for control. For example, a significant drop in ambient temperature would eventually result in far fewer air conditioning loads. System operations would need to take account of such temporal changes in load controllability.

### B. Plug-In Electric Vehicles

PEVs are expected to comprise around 25% of all automobile sales in the United States by 2020 [59]. At those penetration levels, PEVs will account for 3%–6% of total electrical energy consumption. It is anticipated that most vehicles will charge overnight, when other loads are at a minimum. The proportion of PEV load during that period will therefore be quite high. Vehicle charging tends to be rather flexible, though must observe the owner-specified completion time. PEVs therefore offer another excellent end-use class for load control.

Motivated by the control strategy for TCLs developed in [33], a hysteretic form of local control can be used to establish system-level controllability of PEV charging loads. The proposed local control strategy is illustrated in Fig. 8. The nominal SoC profile is defined as the linear path obtained by uniform charging, such that the desired total energy $E_{\text{tot}}$ is delivered to the PEV over the period defined by owner-specified start and finish times. The nominal SoC profile lies at the center of a deadband; for this example, the deadband limits are given by

$$\begin{aligned} \Delta_+(t) &= \text{SoC}(t) + 0.05E_{\text{tot}} \\ \Delta_-(t) &= \text{SoC}(t) - 0.05E_{\text{tot}} \end{aligned} \tag{1}$$

where $\text{SoC}(t)$ is the nominal SoC at time $t$.

When the charger is turned on, the SoC actually increases at a rate that is faster than the nominal profile, so

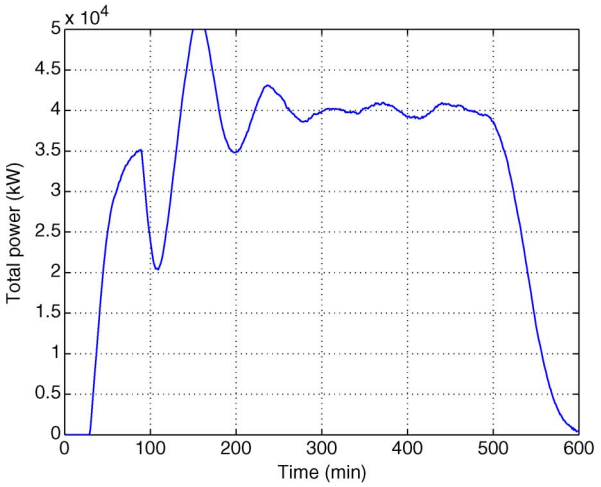


**Fig. 9.** *Total demand of a population of 20 000 PEV loads.*

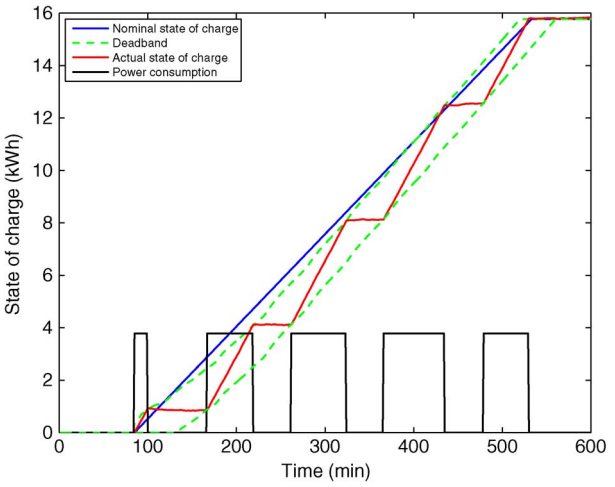


**Fig. 10.** *PEV charging, with control adjustments to the deadband.*

the upper deadband will be encountered. At that point, the charger turns off. The lower deadband steadily increases over time. When the actual SoC intersects the lower deadband, the charger again turns on. This process is repeated until the PEV is fully charged, as shown in Fig. 8. The figure also shows that the charger draws power as a sequence of pulses. Note that if PEVs are capable of modulating their power consumption, the simple hysteretic local controller used here could be replaced by a more sophisticated controller.

This approach is appealing from a system-level control perspective, as the local control objective is a functional quantity (in this case, a desired SoC profile) rather than power. As discussed earlier, the system-level controller can then make adjustments to the end-use function. Provided it is possible to predict changes in aggregated power demand as a function of end-use function adjustments, system-level control decisions can be made on the basis of preserving end-use performance.

To illustrate the system-level consequences of this charging process, we will use an example that consists of 20 000 PEV loads. The charging start and finish times for the PEVs were uniformly distributed over the ranges 30–90 and 510–570 min, respectively. The energy required by each PEV was uniformly distributed over 12–20 kWh, and charger power was uniformly distributed over 3–5 kW. The total load drawn by the collection of PEVs is shown in Fig. 9. Early oscillations are the result of transient synchrony across the population of loads. Over time, however, the total demand reaches steady state as behavior becomes more heterogeneous. Demand ultimately returns to zero as all PEVs progressively complete their charging.

Control of PEV charging load can be achieved by adjusting the SoC deadband, akin to adjusting the temperature deadband in the control methodology developed for TCLs in [33] and outlined in Section VI-A. Referring to (1), the deadband adjustment has the form

$$\begin{aligned}\Delta_+(t) &= \mathrm{SoC}(t) + (0.05 + u(t))E_{\mathrm{tot}} \\ \Delta_-(t) &= \mathrm{SoC}(t) - (0.05 - u(t))E_{\mathrm{tot}}\end{aligned} \tag{2}$$

where $u(t)$ is the control input at time $t$. This is illustrated in Fig. 10. Notice that adjustments to the deadband cause variations in the on/off timing of the charger, but that the owner-specified completion time is still honored. This can be observed by comparing Figs. 8 and 10.

Whilst variations in the duration of the power pulses are inconsequential for individual PEVs, their cumulative effect over a large number of PEVs can be quite significant. Controllability of total PEV demand can be illustrated using a tracking example. Referring to Fig. 11, the controller switches on at 40 min, and forces total PEV demand to follow the path indicated by the dashed line, until 450 min. Beyond that time, the control signal ramps back to zero. A simple integral form of control is used, with the deadband adjustment generated according to

$$u(t) = u(t-1) + K(P_{\mathrm{des}}(t-1) - P_{\mathrm{tot}}(t-1)) \tag{3}$$

where $P_{\mathrm{des}}$ is the power along the desired track, and $P_{\mathrm{tot}}$ is the total power of the PEV population. By lowering the deadband, PEV chargers that were about to turn off do so a little earlier, and chargers that were already off remain off a little longer. The overall effect is a reduction in load. Similar logic applies for raising load. The control signal corresponding to the response of Fig. 11 is shown in Fig. 12. The outcome of this strategy, in terms of its effect

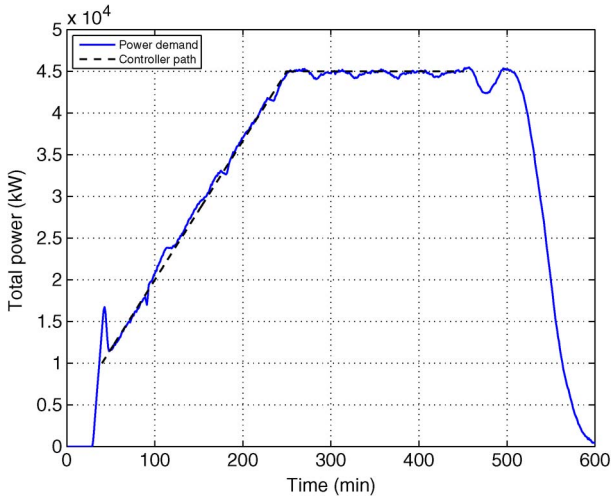


**Fig. 11.** *Total demand of 20 000 PEV loads, controlled to track a specified path.*

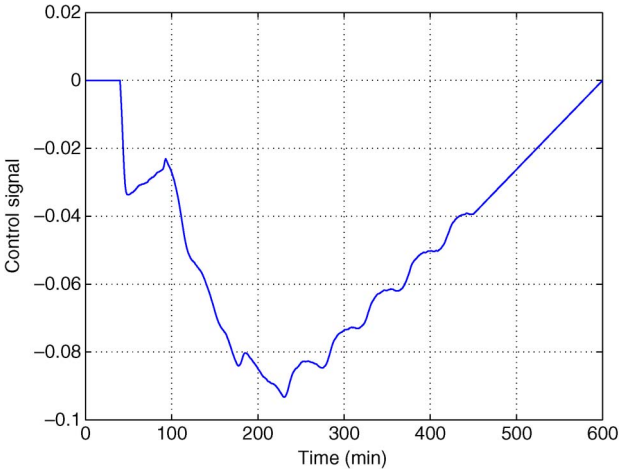


**Fig. 12.** *PEV control signal for driving the total load response of Fig. 11.*

on an individual PEV, is shown in Fig. 10. Notice that significant controllability of total demand can be achieved with quite modest adjustments to PEV deadbands.

The control algorithm (3) adopted for this illustration takes no account of the underlying dynamics of the population of PEVs. A more sophisticated controller tailored to a model of population dynamics could significantly improve tracking fidelity. Furthermore, the control strategy used in the example does not consider the diminishing availability of load control as PEVs progressively reach their fully charged state. As suggested in Section III-D, control strategies must ensure that such temporal constraints are addressed. These issues are the focus of ongoing research.

## VII. SUMMARY AND CONCLUSION

This paper has explored some of the opportunities and challenges associated with implementing fully responsive, nondisruptive control strategies for aggregated electric loads. We discuss these control actions in the specific context of grid operations such as automatic generation control, spinning reserve, and economic dispatch/unit commitment.

The central challenge for nondisruptive load control is that there are dual, often competing, control objectives: first, to achieve desirable aggregated power consumption patterns, and second, to maintain acceptable end-use performance (e.g., temperature of conditioned spaces or PEV battery SoC). These objectives can be managed by quantifying metrics of load availability (a measure of the physical capacity available for control) and willingness to participate in aggregated control activities (determined by constraints on the quality of end-use function). To the extent that total energy consumption over a time horizon determines end-use function (rather than instantaneous power), control decisions at one moment in time will constrain the set of possible future decisions. These temporal constraints are directly related to the willingness and availability measures. It is likely that these three factors—availability, willingness, and temporal constraints—will need to be considered in some form or another for any successful aggregated load control strategy.

Communications infrastructure must be considered in the design of any control paradigm. In the case of aggregated load control, issues of data and infrastructure ownership, latency in communications processes, and the frequency of control signal updates will be particularly important. Because it will likely be ubiquitous, AMI has good potential as a communications platform, but is limited in its ability to collect device-specific information on rapid time scales. Perhaps the most promising way to use AMI will be to broadcast one common signal to all loads (which can be done on relatively short time scales) and then extract feedback control information from aggregate power measurements at distribution substations. The choice of the broadcasted signal is an important consideration; a functional quantity, such as thermostat setpoint deviation, may make the most sense since it leaves the task of maintaining end-use function to the local controller. However, using an input signal other than desired power consumption means that a model which relates the aggregate power to changes in the functional quantity will be needed for any type of forward-looking control strategy.

There are a number of possible control architectures, including centralized, hierarchical, and distributed controllers. The hierarchical model may hold the most promise because: 1) it creates an avenue for third parties to organize loads and bid them into energy and ancillary service markets, and 2) it may allow the system operator to conceptualize load groups managed by each aggregator as individual resources, similar to individual units on the supply side. This provides a relatively seamless route to fit

aggregated load resources into the legacy control paradigm. However, the distributed control approach, where loads provide the equivalent of generator droop, is also promising. ■

## Acknowledgment

The authors would like to thank the three anonymous reviewers for their insightful comments, some of which significantly changed parts of this paper.

## REFERENCES

[1] G. Strbac, "Demand side management: Benefits and challenges," *Energy Policy*, vol. 36, no. 12, pp. 4419–4426, 2008.

[2] B. Kirby, "Spinning reserve from responsive loads," Oak Ridge Nat. Lab., Oak Ridge, TN, Tech. Rep. ORNL/TM-2003/19, Mar. 2003.

[3] M. Klobasa, "Analysis of demand response and wind integration in Germany's electricity market," *IET Renewable Power Generat.*, vol. 4, no. 1, pp. 55–63, 2010.

[4] C. Woo, E. Kollman, R. Orans, S. Price, and B. Horii, "Now that California has AMI, what can the state do with it?" *Energy Policy*, vol. 36, no. 4, pp. 1366–1374, 2008.

[5] I. A. Hiskens and B. Gong, "MPC-based load shedding for voltage stability enhancement," in *Proc. 44th IEEE Conf. Decision Control*, Seville, Spain, Dec. 2005, pp. 4463–4468.

[6] B. Ramanathan and V. Vittal, "Small-disturbance angle stability enhancement through direct load control: Part I—Framework development," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 773–781, May 2006.

[7] A. Wood and B. Wollenberg, *Power Generation Operation and Control*. New York: Wiley, 1996.

[8] W.-C. Chu, B.-K. Chen, and C.-K. Fu, "Scheduling of direct load control to minimize load reduction for a utility suffering from generation shortage," *IEEE Trans. Power Syst.*, vol. 8, no. 4, pp. 1525–1530, Nov. 1993.

[9] C. Kurucz, D. Brandt, and S. Sim, "A linear programming model for reducing system peak through customer load control programs," *IEEE Trans. Power Syst.*, vol. 11, no. 4, pp. 1817–1824, Nov. 1996.

[10] K. Huang, H. Chin, and Y. Huang, "A model reference adaptive control strategy for interruptible load management," *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 683–689, Feb. 2004.

[11] T.-F. Lee, M.-Y. Cho, Y.-C. Hsiao, P.-J. Chao, and F.-M. Fang, "Optimization and implementation of a load control scheduler using relaxed dynamic programming for large air conditioner loads," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 691–702, May 2008.

[12] L. Yao and H.-R. Lu, "A two-way direct control of central air-conditioning load via the internet," *IEEE Trans. Power Delivery*, vol. 24, no. 1, pp. 240–248, Jan. 2009.

[13] Y. Hsu and C. Su, "Dispatch of direct load control using dynamic programming," *IEEE Trans. Power Syst.*, vol. 6, no. 3, pp. 1056–1061, Aug. 1991.

[14] H. Salehfar and A. Patton, "A production costing methodology for evaluation of direct load control," *IEEE Trans. Power Syst.*, vol. 6, no. 1, pp. 278–284, Feb. 1991.

[15] J. Chen, F. Lee, A. Breipohl, and R. Adapa, "Scheduling direct load control to minimize system operation cost," *IEEE Trans. Power Syst.*, vol. 10, no. 4, pp. 1994–2001, Nov. 1995.

[16] K.-H. Ng and G. Sheble, "Direct load control—A profit-based load management using linear programming," *IEEE Trans. Power Syst.*, vol. 13, no. 2, pp. 688–694, May 1998.

[17] K. Huang and Y. Huang, "Integrating direct load control with interruptible load management to provide instantaneous reserves for ancillary services," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1626–1634, Aug. 2004.

[18] L. Goel, Q. Wu, and P. Wang, "Fuzzy logic-based direct load control of air conditioning loads considering nodal reliability characteristics in restructured power systems," *Electr. Power Syst. Res.*, vol. 80, no. 1, pp. 98–107, 2010.

[19] K. Bhattacharyya and M. Crow, "A fuzzy logic based approach to direct load control," *IEEE Trans. Power Syst.*, vol. 11, no. 2, pp. 708–714, May 1996.

[20] B. Ramanathan and V. Vittal, "A framework for evaluation of advanced direct load control with minimum disruption," *IEEE Trans. Power Syst.*, vol. 23, no. 4, pp. 1681–1688, Nov. 2008.

[21] W. Burke and D. Auslander, "Robust control of residential demand response network with low bandwidth input," in *Proc. ASME Dyn. Syst. Control Conf.*, Ann Arbor, MI, Oct. 2008, pp. 413–415.

[22] N. Ruiz, I. Cobelo, and J. Oyarzabal, "A direct load control model for virtual power plant management," *IEEE Trans. Power Syst.*, vol. 24, no. 2, pp. 959–966, May 2009.

[23] Y. Sherif and S. Zahir, "Communication systems for load management," *IEEE Trans. Power Apparatus Syst.*, vol. PAS-104, no. 12, pp. 3329–3337, Dec. 1985.

[24] A. Faruqi, R. Hledik, S. George, J. Bode, P. Mangasarian, I. Rohmund, G. Wikler, D. Ghosh, and S. Yoshida, "A national assessment of demand response potential," Federal Energy Regulatory Commission, Washington, DC, Tech. Rep., 2009.

[25] K. Schisler, T. Sick, and K. Brief, "The role of demand response in ancillary services markets," in *Proc. IEEE/PES Transm. Distrib. Conf. Expo.*, Apr. 2008. DOI: 10.1109/TDC.2008.4517087.

[26] E. Koch and M. Piette, "Architecture concepts and technical issues for an open, interoperable automated demand response infrastructure," Lawrence Berkeley Nat. Lab., Berkeley, CA, Tech. Rep. LBNL-63664, Oct. 2007.

[27] P. Kundur, *Power System Stability and Control*, New York: McGraw-Hill, 1994.

[28] N. Jaleeli, L. VanSlyck, D. Ewart, L. Fink, and A. Hoffmann, "Understanding automatic generation control," *IEEE Trans. Power Syst.*, vol. 7, no. 3, pp. 1106–1122, Aug. 1992.

[29] A. Venkat, I. Hiskens, J. Rawlings, and S. Wright, "Distributed MPC strategies with application to power system automatic generation control," *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 6, pp. 1192–1206, Nov. 2008.

[30] J. Short, D. Infield, and L. Freris, "Stabilization of grid frequency through dynamic demand control," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1284–1293, Aug. 2007.

[31] U.K. Market Transformation Program, "Dynamic demand control of domestic appliances, Market Transformation Programme," Tech. Rep., 2008.

[32] D. Hammerstrom, J. Brous, D. Chassin, G. Horst, R. Kajfasz, P. Michie, T. Oliver, T. Carlon, C. Eustis, O. Jarvegren, W. Marek, R. Munson, and R. Pratt, "Pacific Northwest GridWise Testbed Demonstration Projects, Part II. Grid Friendly Appliance Project," Pacific Northwest Nat. Lab., Richland, WA, Tech. Rep. PNNL-17079, Oct. 2007.

[33] D. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy," *Energy Conv. Manage.*, vol. 50, no. 9, pp. 1389–1400, May 2009.

[34] M. Piette, S. Kiliccote, and G. Ghatiker, "Linking continuous energy management and open automated demand response," Lawrence Berkeley Nat. Lab., Berkeley, CA, Tech. Rep. LBNL-1361E, Nov. 2008.

[35] P. Cappers, C. Goldman, and D. Kathan, "Demand response in U.S. electricity markets: Empirical evidence," *Energy*, vol. 35, no. 4, pp. 1526–1535, 2010.

[36] J. Zarnikau, "Demand participation in the restructured Electric Reliability Council of Texas market," *Energy*, vol. 35, no. 4, pp. 1536–1543, 2010.

[37] S. Isser, "FERC order 719 and demand response in ISO markets," Good Company Associates/ERCOT, Tech. Rep., 2009.

[38] Electric Reliability Council of Texas Inc., "Controllable load qualification: ERCOT process for qualification testing of controllable loads in accordance with ERCOT protocols and guides," Austin, TX, May 2007.

[39] P. Wattles, personal communication, Aug. 2009, Electric Reliability Council of Texas.

[40] J. Eto, J. Nelson-Hoffman, C. Torres, S. Hirth, B. Yinger, J. Kueck, B. Kirby, C. Bernier, R. Wright, A. Barat, and D. Watson, "Demand response spinning reserve demonstration," Lawrence Berkeley Nat. Lab., Berkeley, CA, Tech. Rep. LBNL-62761, May 2007.

[41] G. Heffner, C. Goldman, B. Kirby, and M. Kintner-Meyer, "Loads providing ancillary services: Review of international experience," Lawrence Berkeley Nat. Lab., Berkeley, CA, Tech. Rep. LBNL-62701, May 2007.

[42] K. Spees and L. Lave, "Demand response and electricity market efficiency," *Electricity J.*, vol. 20, no. 3, pp. 69–85, 2007.

[43] E. Woychik, "Optimizing demand response," *Public Utilities Fortnightly*, vol. 146, no. 5, pp. 52–56, May 2008.

[44] C. Goldman, N. Hopper, R. Bharvirkar, B. Neenan, and P. Cappers, "Estimating demand response market potential among large commercial and industrial customers: A scoping study," Lawrence Berkeley Nat. Lab., Berkeley, CA, Tech. Rep., 2007.

[45] R. Belhomme, R. Cerero Real De Asua, G. Valtorta, A. Paice, F. Bouffard, R. Rooth, and A. Losi, "ADDRESS—Active demand for the smart grids of the future," presented at

the CIRED Seminar 2008: GmartGrids for Distribution, Frankfurt, Germany, Jun. 2008, Paper 0080.

[46] S. Koch, D. Meier, M. Zima, M. Wiederkehr, and G. Andersson, "An active coordination approach for thermal household appliances—Local communication and calculation tasks in the household," in *Proc. IEEE PowerTech*, Bucharest, Romania, Jun. 2009. DOI: 10.1109/PTC.2009.5281787.

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

[48] S. Borenstein, M. Jaske, and A. Rosenfeld, "Dynamic pricing, advanced metering, and demand response in electricity markets," Center for the Study of Energy Markets, Tech. Rep. CSEMWP105, Oct. 2002.

[49] D. Kirschen, "Demand-side view of electricity markets," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 520–527, May 2003.

[50] U.S. Department of Energy, "Benefits of demand response in electricity markets and recommendations for achieving them," U.S. Dept. Energy, Washington, DC, Tech. Rep., Feb. 2006.

[51] H. Chao, S. Oren, S. Smith, and R. Wislon, "Multilevel demand subscription pricing for electric power," *Energy Econom.*, vol. 8, no. 4, pp. 199–217, Oct. 1986.

[52] S. Oren and S. Smith, "Design and management of curtailable electricity service to reduce annual peaks," *Oper. Res.*, vol. 40, no. 2, pp. 213–228, 1992.

[53] J. Bushnell, B. Hobbs, and F. Wolak, "When it comes to demand response, is FERC its own worst enemy?" *Electricity J.*, vol. 22, no. 8, pp. 9–18, 2009.

[54] J. Mathieu, A. Gadgil, D. Callaway, P. Price, and S. Kiliccote, "Characterizing the response of commercial and industrial facilities to dynamic pricing signals from the utility," in *Proc. ASME 4th Int. Conf. Energy Sustainability*, 2010.

[55] A. Kashyap and D. Callaway, "Estimating the probability of load curtailment in power systems with responsive distributed storage," in *Proc. 11th Int. Conf. Probab. Methods Appl. Power Syst.*, 2010, pp. 18–23.

[56] M. Huang, P. Caines, and R. Malhame, "Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized $\varepsilon$-Nash equilibria," *IEEE Trans. Autom. Control*, vol. 52, no. 9, pp. 1560–1571, Sep. 2007.

[57] G. Goodwin, S. Graebe, and M. Salgado, *Control System Design*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[58] Energy Information Administration, "Residential energy consumption survey," U.S. Dept. Energy, Washington, DC, Tech. Rep., 2001.

[59] S. Hadley, "Impact of plug-in hybrid vehicles on the electric grid," Oak Ridge Nat. Lab., Oak Ridge, TN, Tech. Rep. ORNL/TM-2006/554, Oct. 2006.

## ABOUT THE AUTHORS

**Duncan S. Callaway** (Member, IEEE) received the B.S. degree in mechanical engineering from the University of Rochester, Rochester, NY, in 1995 and the Ph.D. degree in theoretical and applied mechanics from Cornell University, Ithaca NY, in 2001.

Currently, he is an Assistant Professor of Energy and Resources and Mechanical Engineering at the University of California, Berkeley. Prior to joining the University of California, he was first a National Science Foundation (NSF) Postdoctoral Fellow at the Department of Environmental Science and Policy, University of California, Davis, subsequently worked as a Senior Engineer at Davis Energy Group, Davis, CA, and PowerLight Corporation, Berkeley CA, and was most recently a Research Scientist at the University of Michigan, Ann Arbor. His current research interests are in the areas of power management, modeling and control of aggregated storage devices, spatially distributed energy resources, and environmental impact assessment of energy technologies.

**Ian A. Hiskens** (Fellow, IEEE) received the B.Eng. degree in electrical engineering and the B.App.Sc. degree in mathematics from the Capricornia Institute of Advanced Education, Rockhampton, Australia, in 1980 and 1983, respectively and the Ph.D. degree in electrical engineering from the University of Newcastle, Australia, in 1991.

Currently, he is the Vennema Professor of Engineering at the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. He has held prior appointments in the Queensland electricity supply industry (for ten years), and various universities in Australia and the United States. His major research interests lie in the area of power system analysis, in particular system dynamics and control, and security assessment. His recent activity has focused largely on integration of new forms of generation and load. Other research interests include nonlinear and hybrid dynamical systems.

Dr. Hiskens is actively involved in various IEEE societies, and is Treasurer of the IEEE Systems Council. He is a Fellow of Engineers Australia and a Chartered Professional Engineer in Australia.