Distributed Peak Shaving for Small Aggregations of Cyclic Loads

Kevin J. Kircher, *Member, IEEE,* Adedayo O. Aderibole, Leslie K. Norford, and Steven B. Leeb, *Fellow, IEEE*

***Abstract*—Analogous to the way a good driver is aware of neighboring cars, electrical loads can coordinate with other loads within a building or section of a distribution grid. This paper develops methods that enable groups of cyclic loads (devices that turn on and off periodically to maintain setpoints) to reduce their peak aggregate power demand. The methods accommodate a wide variety of cyclic loads, including those with nonlinear or unknown dynamics, and can be implemented in a fully distributed fashion. This paper targets settings with a few hundred cyclic loads or fewer, where the methods developed here could reduce demand peaks significantly while maintaining or improving quality of service. This could save ratepayers money on monthly demand charges, decrease fuel use in microgrids, or extend the life of power delivery equipment.**

***Index Terms*—Demand Response, Peak Shaving, Cyclic Loads, Thermostatically-Controlled Loads**

Nomenclature

L	Number of loads.
ℓ	Load index.
$\mathcal{L}$	Set of load indices.
K	Number of discrete time steps.
k	Time index.
$\mathcal{K}$	Set of time indices.
Δt	Time step duration (h).
$P_\ell(k)$	Electric power (kW).
$\overline{P}_\ell$	Electric power capacity (kW).
P^{cap}	Aggregate power cap (kW).
d_ℓ	Duty cycle.
$u_\ell(k)$	On/off control input, in $\{0, 1\}$.
$y_\ell(k)$	Measured output, normalized to $[0, 1]$.
$\mathcal{A}(k)$	Set of load indices dispatched under conventional bang-bang control.
$s_\ell(k)$	Priority score.
$s^\star(k)$	Priority score threshold.
$x_\ell(k)$	Consecutive on-time (h).
$T_\ell(k)$	Indoor temperature (°C).
$T_\ell^{\text{set}}(k)$	Indoor temperature setpoint (°C).
$\theta_\ell(k)$	Outdoor temperature (°C).
$\dot{Q}_\ell(k)$	Exogenous thermal power (kW).
R_ℓ	Thermal resistance (°C/kW).
C_ℓ	Thermal capacitance (kWh/°C).
η_ℓ	Coefficient of performance.
a_ℓ	Discrete-time dynamics parameter.
ε_ℓ	Allowed deviation from setpoint (°C).
T_ℓ^{des}	Design indoor temperature (°C).
θ_ℓ^{des}	Design outdoor temperature (°C).
$\dot{Q}_\ell^{\text{des}}$	Design exogenous thermal power (kW).
P_ℓ^{des}	Design electric power (kW).
ρ_ℓ	Oversize ratio relative to design load.

K.J. Kircher (kircher@mit.edu), A.O. Aderibole and S.B. Leeb are with the Department of Electrical Engineering and Computer Science and L.K. Norford is with the Department of Architecture, all at the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

I. Demand Peaks and Cyclic Loads

PEAKS in electricity demand shape power system operations at various scales. In bulk transmission grids, demand peaks drive generation capacity expansion and increase the use of inefficient peaking plants. This paper focuses on smaller-scale applications, such as a large apartment building or hotel, a microgrid, or a section of a distribution grid covering a neighborhood or city block. At these scales, there are several motivations to reduce peak demand. In large buildings, fees based on monthly demand peaks can comprise 30–70% of electricity bills [1]. In microgrids, demand peaks can increase fuel use by triggering dispatch of backup generators, or by causing generators to operate at higher load and lower efficiency [2]. In sections of distribution grids, demand peaks can stress transformers and power lines, increasing the risk of equipment failure [3].

To reduce demand peaks at these scales, this paper develops control techniques for aggregations of cyclic loads, meaning fixed-speed electrical devices that turn on and off periodically to regulate a measured variable within a band centered on a setpoint. Specific demonstrations are made for thermostatically-controlled loads (TCLs) – cyclic loads that regulate temperature – as TCLs such as air conditioners and heat pumps are key drivers of demand peaks. However, the control techniques developed here also apply to a variety of other cyclic loads, such as fixed-speed pumps, fans, compressors, dehumidifiers, and vacuum systems.

Fig. 1 illustrates the operating cycle of a thermostatically-controlled air conditioner. The air conditioner is initially off, so the indoor air warms until it reaches the thermostat's

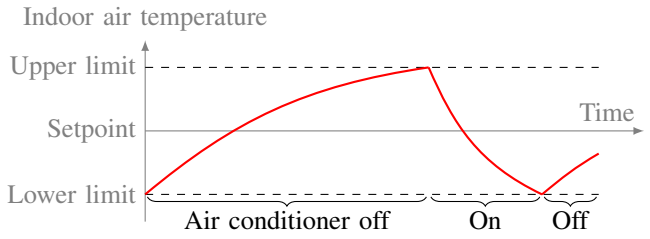

Fig. 1. A thermostatically-controlled air conditioner cycles off and on to keep the measured indoor temperature (red) within a band centered on a setpoint.

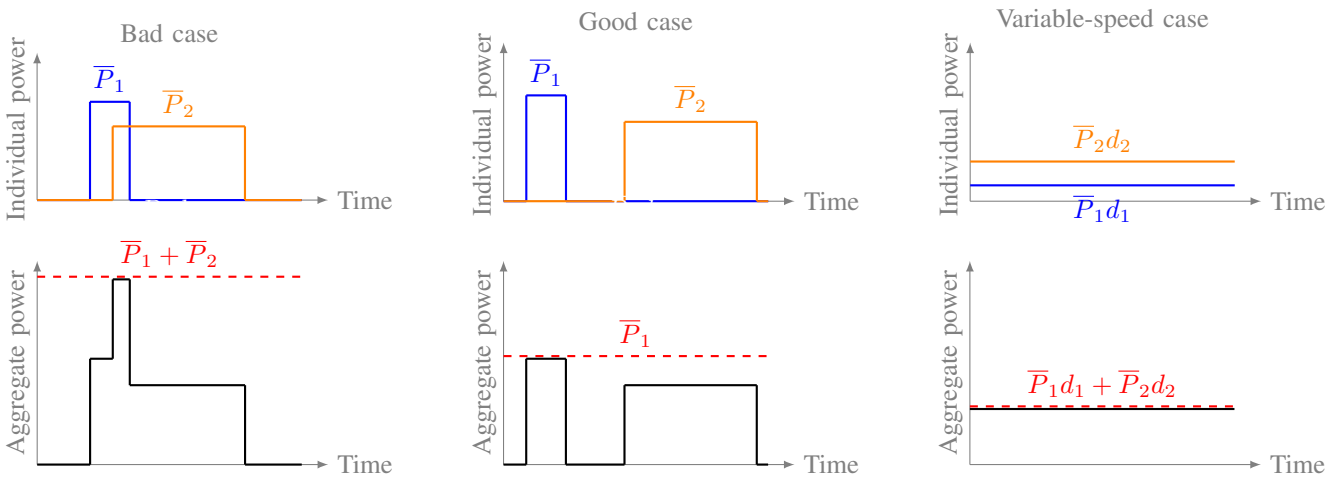

Fig. 2. Individual (top) and aggregate (bottom) power of two cyclic loads. At left, the loads are briefly on at the same time, but in the center column they coordinate to stagger their operation. The variable-speed case (right) lower-bounds the peak aggregate demand (dashed red lines) that can be achieved.

upper temperature limit. The air conditioner then switches on, driving the temperature down to the lower limit, at which point the air conditioner switches off and the process repeats.

TCLs have played an important role in demand-side management efforts since the energy price shocks of the 1970s [4]. Early approaches to TCL peak shaving involved utilities turning TCLs off during system peaks, using either a direct load control signal or a phone call to the TCL operator. Later approaches involved subtler methods, such as raising air conditioner temperature setpoints during afternoon peaks.

While these approaches are simple and effective, they may cause significant inconvenience or discomfort to users. For example, turning air conditioners off or raising their temperature setpoints can make buildings uncomfortably hot, humid or stuffy. This paper investigates an alternative approach that does not sacrifice quality of service. This non-intrusive approach is based on the observation that demand peaks in cyclic load aggregations are driven in part by an unfortunate and avoidable coincidence: occasionally, an unnecessarily large fraction of the loads will happen to run at the same time. Loads can avoid this coincidence by coordinating to interleave their operation. This coordination can shave peaks substantially while maintaining quality of service.

Fig. 2 illustrates this unfortunate coincidence and the opportunity to avoid it. In the left column of Fig. 2, two cyclic loads are briefly on at the same time, as can be seen from the overlapping individual power profiles in the top left plot. This coincidence causes a peak aggregate demand (the dashed red line in the bottom left plot) of $\overline{P}_1 + \overline{P}_2$, where $\overline{P}_\ell$ (kW) is the electric power capacity of load ℓ. In the center column of Fig. 2, by contrast, the loads coordinate to avoid being on at the same time, as can be seen from the staggering of the individual power profiles in the top center plot. This coordination reduces the peak aggregate demand (the dashed red line in the bottom center plot) to $\overline{P}_1$. In Sec. V, this paper develops methods to coordinate arbitrary numbers of loads for similar peak reductions.

The right-hand column of Fig. 2 illustrates the operation of two hypothetical variable-speed loads that could replace the cyclic loads. While each cyclic load alternates between turning off and running at full capacity, variable-speed load ℓ would run continuously at partial capacity $\overline{P}_\ell d_\ell$, as can be seen in the top right plot. Here d_ℓ is the duty cycle (the ratio of 'on' time to total operating time) of cyclic load ℓ. The bottom row of plots in Fig. 2 shows that the aggregate power of the variable-speed loads, $\overline{P}_1 d_1 + \overline{P}_2 d_2$, is a lower bound on the peak demand that can be achieved by coordinating the two cyclic loads. In Sec. IV, this paper establishes that this lower bound is not specific to the simple two-load example in Fig. 2, but in fact provides a fundamental performance bound for any number of cyclic loads and any coordination method. The variable-speed bound illustrated in Fig. 2 is therefore a valuable benchmark for practical control systems.

In Sec. VI, this paper simulates TCL coordination using the methods developed in Sec. V. In these simulations, the methods reduce peak demand to the variable-speed bound under a wide range of boundary conditions, load characteristics and device parameters. This suggests that the control methods are essentially optimal for the problem considered here: they reduce peak demand as much as possible without compromising quality of service. Simulations also show how the potential peak reduction scales with the number of loads. The potential is significant for aggregations of a few hundred loads or fewer.

II. Literature Review and Contributions

There is an extensive literature on providing grid services from cyclic loads, particularly TCLs. For example, research has shown that TCL aggregations can arbitrage dynamic energy prices [5]–[8], track aggregate power reference signals for frequency regulation or renewable supply following [8]–[16], and reduce peak demand [16]–[28]. This literature review focuses on the peak shaving studies [16]–[28], as they are most relevant to the current scope. While shifting load based on wholesale prices can indirectly reduce peak system demand,

wholesale price spikes may not coincide with local demand peaks at the small scales considered here.

Within the TCL peak shaving literature, most studies control the loads' on/off states directly [17]–[24], [26], and this paper uses the same approach. By contrast, [16] and [25] control on/off states indirectly by adjusting temperature setpoints and supply voltages, respectively. To decide control actions, some studies use numerical optimization, typically solving linear or mixed-integer linear programs [16], [17], [20], [23], [27]. Optimization problems are solved statically [17], [20], [23] or in a receding-horizon fashion [16], [27]. Another group of studies uses scheduling heuristics that prioritize loads based on a variety of scoring metrics [18], [19], [21], [22], [26]. This paper generalizes these heuristics to accommodate arbitrary priority-scoring metrics, as well as hard constraints such as maximum switching frequencies or minimum on- or off-times. This paper also shows how priority-based control can be implemented in a fully distributed fashion. The methods here have conceptual similarities to the market-based approach in [24], where loads submit price and quantity bids to a central agent that clears the market.

Most TCL peak shaving studies assume that TCLs have low-order linear time-invariant (LTI) dynamics [16], [17], [20]–[27]. While this assumption simplifies control design and may be realistic in some situations, it is not always a good approximation. In [29], for example, Xu *et al.* find significant inaccuracies in both first- and second-order LTI water heater models. Model inaccuracies may arise from lumped approximations to continuous temperature distributions [30], from nonlinearities due to thermal radiation or temperature-dependent convection coefficients [31], or from the time-variation of device efficiencies that depend nontrivially on ambient temperatures. By avoiding restrictive modeling assumptions, model-free approaches such as [19] can accommodate a wider class of loads, including non-thermal loads and loads with high-order, nonlinear, time-varying, or unknown dynamics. Model-free approaches also avoid the need for physics-based modeling or system identification and are not vulnerable to structural or parametric model errors.

Most TCL peak shaving studies also propose central control architectures, where one agent makes control decisions for all loads [16]–[26]. This approach carries significant cybersecurity risks, particularly vulnerability to denial-of-service attacks on the central agent [27]. By avoiding dependence on a single point of failure (the central agent), fully distributed approaches such as [27] can improve robustness. This paper develops a peak shaving control framework for cyclic loads that is both model-free and fully distributed. To the best of the authors' knowledge, this is the first such framework in the literature.

In summary, this paper makes three main contributions. **First**, it develops model-free, fully distributed peak shaving control methods for a wide class of fixed-speed loads. In these methods, each load quantifies the urgency of its energy needs via a priority score, which it shares with its neighbors. Lower-priority loads turn off to make room for higher-priority loads. The control methods have the additional advantages of requiring little communication or computation and no information about user behavior. **Second**, this paper establishes a new and fundamental bound on the peak demand that any coordination method can achieve without sacrificing quality of service. This bound is a valuable benchmark for practical control systems. **Third**, this paper simulates peak shaving control of heterogeneous aggregations of air conditioners during California's extreme heat wave of mid-August, 2020, which caused rolling blackouts throughout the state. In these simulations, the control methods developed here achieve the peak shaving bound while maintaining or improving quality of service. In addition to building confidence in the proposed methods, these simulations illustrate how the achievable peak reduction scales with the number of loads and their degree of oversizing. In particular, this paper shows that non-intrusive peak shaving is fundamentally a small-aggregation problem, a result not found in existing literature.

III. Mathematical Framework

The mathematical framework in this paper involves a set $\mathcal{L} = \{1, \dots, L\}$ of cyclic loads and a discrete time span $\mathcal{K} = \{1, \dots, K\}$. At each time $k \in \mathcal{K}$, each load $\ell \in \mathcal{L}$ observes a local output $y_\ell(k)$ and potentially receives additional information from other loads. Each load ℓ then decides the control input

$$u_\ell(k) \in \{0, 1\}, \tag{1}$$

with 0 meaning 'off' and 1 meaning 'on'. Given the control inputs $u_1(k)$, …, $u_L(k)$ and the system state at time k, the system evolves to a new state at time $k+1$. This evolution may involve exogenous inputs related, *e.g.*, to weather or human behavior. Load ℓ then measures $y_\ell(k+1)$ and the process repeats. The constraint

$$0 \le y_\ell(k) \le 1 \tag{2}$$

encodes the requirement that the measured output remain within a given band. The input and output are defined so that, all else being equal, $y_\ell(k+1)$ will be larger if $u_\ell(k) = 1$ than if $u_\ell(k) = 0$. Beyond these assumptions, the system dynamics are arbitrary. They could be nonlinear, high-dimensional, stochastic, time-varying, coupled across loads, *etc*. The controller requires no additional measurements beyond $y_\ell(k)$, which conventional controllers also require.

For example, a pump might turn on and off periodically to maintain the measured water level $h(k)$ (m) in a reservoir between minimum and maximum levels, $\underline{h}$ and $\overline{h}$ (m), despite water withdrawals. This pump defines the output $y_\ell(k) = (h(k) - \underline{h})/(\overline{h} - \underline{h})$, so that (2) holds if and only if $\underline{h} \le h(k) \le \overline{h}$. All else being equal, the scaled water level $y_\ell(k+1)$ will be larger if the pump is on ($u_\ell(k) = 1$) than off ($u_\ell(k) = 0$), as the framework requires.

IV. Non-Intrusive Peak Shaving Bound

To satisfy its operational constraints (2) under given boundary conditions, each load ℓ requires a certain duty cycle $d_\ell \in [0, 1]$ over the time span $\mathcal{K}$ [32]. More precisely, the control inputs $u_\ell(1)$, …, $u_\ell(K)$ must satisfy

$$\frac{\Delta t \sum_{k \in \mathcal{K}} u_\ell(k)}{K \Delta t} = \frac{\text{'on' time}}{\text{total time}} = d_\ell,$$

where Δt (h) is the time step duration. In terms of the device powers $P_\ell(k) = u_\ell(k)\overline{P}$, this requirement can be stated as

$$\sum_{k\in\mathcal{K}} P_\ell(k) = K\overline{P}_\ell d_\ell. \tag{3}$$

The problem of minimizing peak aggregate demand subject to the on/off power constraint (1) and the duty cycle constraint (3) can therefore be stated as

$$\begin{array}{ll} \text{minimize} & \max_{k\in\mathcal{K}} \sum_{\ell\in\mathcal{L}} P_\ell(k) \\ \text{subject to} & \sum_{k\in\mathcal{K}} P_\ell(k) = K\overline{P}_\ell d_\ell,\ \ell\in\mathcal{L} \\ & P_\ell(k) \in \left\{0, \overline{P}_\ell\right\},\ k\in\mathcal{K},\ \ell\in\mathcal{L}. \end{array} \tag{4}$$

In (4), the decision variables are the device powers $P_\ell(k)$. The problem data are the duty cycles d_ℓ and device capacities $\overline{P}_\ell$.

Appendix A establishes a lower bound on the optimal value of (4):

$$\max_{k\in\mathcal{K}} \sum_{\ell\in\mathcal{L}} P^\star_\ell(k) \geq \sum_{\ell\in\mathcal{L}} \overline{P}_\ell d_\ell, \tag{5}$$

where $P^\star_\ell(k)$ (kW) denotes any optimal solution to (4). The left-hand side of (5) is the peak demand under any control scheme that is feasible and optimal for (4). The right-hand side of (5) can be interpreted as the peak demand of a fleet of ideal[1] variable-speed loads that hypothetically replace the cyclic loads, with variable-speed load ℓ operating at constant power $\overline{P}_\ell d_\ell$.

The bound (5) provides a fundamental limit on the non-intrusive peak shaving that any method for coordinating cyclic loads can achieve. If peak demand is reduced below this limit, then not all setpoints can be maintained.

Analogies between fixed- and variable-speed loads have provided useful insights elsewhere in the literature. In Lemma 1 of [32], for example, Mahdavi *et al.* show that a single TCL's duty cycle is approximated within a certain error by the part-load ratio of an ideal variable-speed load. While this lemma and the variable-speed bound (5) both draw an analogy between on/off and variable-speed devices, the contexts and conclusions are quite different. The context here is a load aggregation, rather than one load; the conclusion here is an exact lower bound on the achievable peak demand, rather than an approximation of the duty cycle.

V. Control Design

A control system that limits the peak demand of cyclic loads should satisfy the on/off constraint (1) and the output constraint (2). It should also limit peak demand:

$$\max_{k\in\mathcal{K}} \sum_{\ell\in\mathcal{L}} u_\ell(k)\overline{P}_\ell \leq P^{\text{cap}}, \tag{6}$$

where P^{cap} (kW) is a given power cap. In practice, P^{cap} could be set based on historical data or adaptively tuned. It is not always possible to satisfy both (2) and (6); when a conflict occurs, the output constraint (2) should take precedence. To reduce wear and tear on devices, the control system should also limit the fleet-average device switching frequency,

$$\frac{1}{KL\Delta t} \sum_{\ell\in\mathcal{L}} \sum_{k\in\mathcal{K}} \left| u_\ell(k) - u_\ell(k-1) \right|. \tag{7}$$

[1]The variable-speed loads in this analogy are ideal in that they perfectly maintain their setpoints under steady conditions, and use the same amount of energy as the corresponding fixed-speed loads. Real variable-speed loads typically use less energy than fixed-speed loads, as efficiencies are higher at part load than at full load.

A. Conventional Bang-Bang Control

Conventionally, most cyclic loads use a form of bang-bang control. In this approach, each load's controller independently keeps its output in the acceptable range with no communication. The resulting control law is

$$u_\ell(k) = \begin{cases} 0 & \text{if } y_\ell(k) > 1 \\ 1 & \text{if } y_\ell(k) < 0 \\ u_\ell(k-1) & \text{otherwise.} \end{cases}$$

A thermostatic air conditioner, for example, switches on if its load is too hot ($y_\ell(k) < 0$), switches off if its load is too cold ($y_\ell(k) > 1$), and otherwise maintains its previous on/off state.

It will be convenient to write the bang-bang control law as

$$u_\ell(k) = \begin{cases} 1 & \text{if } (y_\ell(k), u_\ell(k-1)) \in \mathcal{A}_\ell(k) \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\mathcal{A}_\ell(k) = \left\{ (y,u) \in \mathbf{R}^2 \;\middle|\; \begin{array}{l} y < 0 \text{ or} \\ 0 \leq y \leq 1 \text{ and } u = 1 \end{array} \right\}.$$

It will also be convenient to write the set of loads dispatched under bang-bang control as

$$\mathcal{A}(k) = \left\{ \ell \in \mathcal{L} \mid (y_\ell(k), u_\ell(k-1)) \in \mathcal{A}_\ell(k) \right\}.$$

B. Priority Control

This section proposes a family of priority-based control methods aimed at satisfying the peak demand constraint (6). In the proposed methods, each load locally computes a priority score $s_\ell(k)$. Only those loads in $\mathcal{A}(k)$ whose priority scores exceed a threshold $s^\star(k)$ are dispatched:

$$u_\ell(k) = \begin{cases} 1 & \begin{array}{l}\text{if } (y_\ell(k), u_\ell(k-1)) \in \mathcal{A}_\ell(k) \\ \text{and } s_\ell(k) \geq s^\star(k)\end{array} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Constraints are accommodated by allowing the priority scores to take on infinite values. If load ℓ must (or must not) run, for example to respect a short-cycling constraint, then it sets $s_\ell(k) = \infty$ (or $-\infty$). This ensures that $s_\ell(k) \geq s^\star(k)$ (or $<$) and that load ℓ is dispatched (or not dispatched).

Algorithm 1 Priority threshold.

input: P^{cap}, $\mathcal{A}(k)$; $\overline{P}_\ell$ and $s_\ell(k)$ for all $\ell \in \mathcal{A}(k)$
  initialize $\mathcal{A} = \mathcal{A}(k)$
  while $\mathcal{A} \neq \emptyset$, $\min\{s_\ell(k) \mid \ell \in \mathcal{A}\} < \infty$, and $\sum_{\ell\in\mathcal{A}} \overline{P}_\ell > P^{\text{cap}}$
    remove an $\ell' \in \operatorname{argmin}\{s_\ell(k) \mid \ell \in \mathcal{A}\}$ from $\mathcal{A}$
  end while
  return $s^\star(k) = \min\{s_\ell(k) \mid \ell \in \mathcal{A}\}$

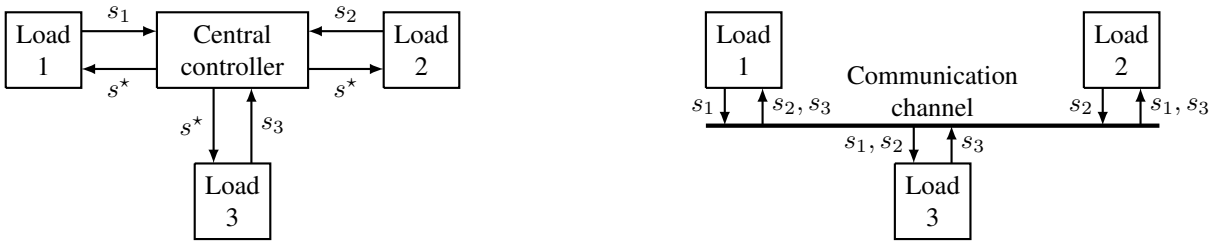

Fig. 3. Possible communication architectures for centralized (left) and distributed (right) control in a three-load network.

Algorithm 1 computes the priority threshold $s^\star(k)$ by iteratively pruning the lowest-priority elements from the set $\mathcal{A}(k)$. It proceeds until one of three conditions are met: (a) the set is empty, meaning all loads that would run under conventional bang-bang control are dispatched; (b) all loads remaining in the set have infinite priority, meaning they must be dispatched; or (c) the aggregate power falls below the demand cap.

An important component of priority control is the metric used to score each load's priority. To satisfy the output constraint (2), a natural choice is $s_\ell(k) = -y_\ell(k)$. This metric prioritizes loads with smaller outputs, meaning an air conditioner has higher priority if its space temperature is higher. Other scoring metrics are possible; for example, [13] and [19] prioritize TCLs by the time remaining until their temperatures exit their thermostatic bands. Another possible scoring metric is

$$s_\ell(k) = -x_\ell(k), \tag{9}$$

where $x_\ell(k)$ is the *on-time* of load ℓ, *i.e.*, the duration that load ℓ has been consecutively running. The on-time dynamics are

$$x_\ell(k+1) = \begin{cases} x_\ell(k) + \Delta t & \text{if } u_\ell(k) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The on-time $x_\ell(k)$ can be viewed as a surrogate for the normalized output $y_\ell(k)$. For example, air conditioners that have been on longer (larger $x_\ell(k)$) tend to have colder space temperatures (larger $y_\ell(k)$). On-time is potentially of interest as a scoring metric if the controller has access to the loads' on/off states, but not to the internal measurements required to compute the outputs $y_\ell(k)$. Scoring via on-time can also reduce communication requirements, as discussed in Sec. V-D.

C. Distributed Implementation

Priority control is straightforward to implement centrally, as illustrated at left in Fig. 3. At each time k, loads $\ell \notin \mathcal{A}(k)$ turn off (or remain off). Loads $\ell \in \mathcal{A}(k)$ report their priority scores $s_\ell(k)$ to the central controller. The central controller runs Algorithm 1 and broadcasts the priority threshold $s^\star(k)$. Loads $\ell \in \mathcal{A}(k)$ compare their scores to the threshold and turn on (or remain on) only if $s_\ell(k) \geq s^\star(k)$. Loads $\ell \in \mathcal{A}(k)$ with $s_\ell(k) < s^\star(k)$ turn off (or remain off). If multiple loads have $s_\ell(k) = s^\star(k)$, various tie-break rules can be used.

Priority control can also be implemented in a fully distributed fashion, with no dependence on a central agent, as illustrated at right in Fig. 3. Executing the priority control law (8) at load ℓ requires only the measured output $y_\ell(k)$, the previous control input $u_\ell(k-1)$, and the priority threshold $s^\star(k)$. The measured output and previous control input are local information, but computing the priority threshold $s^\star(k)$ requires nonlocal information: the capacity $\overline{P}_\ell$ and priority score $s_\ell(k)$ for each load $\ell \in \mathcal{A}(k)$. The capacities can be communicated one time, for example when devices join the network. The priority scores must be communicated online, however. Each load can then run Algorithm 1 locally and set its on/off state according to (8).

In summary, a fully distributed implementation of priority control is the following:

1) Each load $\ell \notin \mathcal{A}(k)$ turns off (or remains off).
2) Each load $\ell \in \mathcal{A}(k)$ broadcasts its priority score.
3) Each load $\ell \in \mathcal{A}(k)$ runs Algorithm 1 locally to compute the priority threshold $s^\star(k)$.
4) Each load $\ell \in \mathcal{A}(k)$ executes the control law (8).

Step 3 involves some redundant computation. In the central implementation, the central controller executes Step 3 once, then broadcasts $s^\star(k)$ to all loads. In the distributed implementation, by contrast, all loads $\ell \in \mathcal{A}(k)$ run Step 3 simultaneously. The associated loss of efficiency is small, as the computation involved in Algorithm 1 is essentially sorting a vector of length L, which takes only $O(L \log L)$ flops. These computing requirements are much lower than optimization-based approaches. With $L = 100$, for example, Algorithm 1 runs in 0.0003 s in Matlab on a 2.7 GHz processor. This likely permits implementation on a low-cost microcontroller.

D. Communication Requirements

The data required to implement priority control using Algorithm 1 include the demand cap P^{cap} and the device capacities $\overline{P}_1, \ldots, \overline{P}_L$, which can be communicated offline. In real time, only the priority scores $s_\ell(k)$ for each $\ell \in \mathcal{A}(k)$ must be communicated. (The set $\mathcal{A}(k)$ does not need to be communicated, as $\mathcal{A}(k)$ is just the set of loads that broadcast their priority scores.) In the worst case, at most L floating-point numbers need to be communicated at each time step. These communication requirements are low: with 50 loads, a one-minute time step, and 16-bit floating-point representation, the worst-case network-wide data rate is 13.3 bits per second (less than one bit per second per load).

Communication requirements can be reduced if the on-time metric (9) is used for priority scoring. In this case, loads do not need to broadcast their priority scores at each time step, but only to broadcast notifications when they switch off and on. Given these notifications, each load can keep track of all other

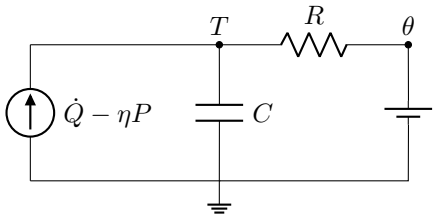

Fig. 4. A 1R1C thermal circuit. The temperature θ and thermal power $\dot{Q}-\eta P$ play the roles of voltage and current sources, respectively.

loads' on/off states and on-times. At each time step, the only input to Algorithm 1 that is not known locally by every load is the set $\mathcal{A}(k)$, which contains the loads that would be on under conventional bang-bang control. Therefore, it suffices for each load $\ell \in \mathcal{A}(k)$ to broadcast a notification that it 'wants to turn on'. Given these notifications, all loads can run Algorithm 1 locally, then execute the priority control law (8).

How much communication this saves depends on how often loads cycle. In the worst case, where each load switches at every time step, L switching notifications are broadcast at each time step. At most L 'I want to turn on' notifications are also broadcast, for a total of $2L$ notifications. Assuming two-bit representation for the three notifications (*e.g.*, `00` for 'I turned off', `01` for 'I turned on', and `10` for 'I want to turn on'), the worst-case data rate is $4L/\Delta t$. In practice, however, loads typically switch a few times per hour and only a small fraction of loads 'want to turn on' at any given time. With 50 devices, a one-minute time step, six switches per device per hour, and $|\mathcal{A}(k)| = 15$ at each time step, the network-wide data rate is 0.67 bits per second.

This data rate is likely low enough for reliable power-line communication [33], [34]. In fact, recent work shows that load switches can be detected by observing the power-line voltage waveform at high sampling rates [35]. This could eliminate the need to broadcast 'I turned off' and 'I turned on' notifications in on-time priority control; the switches themselves would act as notifications. This could introduce occasional errors in switch detection, however. More generally, relying on the power line (or any lossy channel) would risk communication errors. Work may be needed to ensure robustness to these errors in priority control. A simple method to ensure graceful failure is for a load to revert to conventional bang-bang control if its communication reliability degrades, but more sophisticated methods might perform better.

VI. Control Demonstrations

This section simulates a heterogeneous fleet of air conditioners under a variety of control methods. These simulations use a standard TCL model from the literature [16], [20]–[24], [26], [27]. The TCL model serves illustration purposes only. The priority control algorithms simulated here do not have access to the model structure or parameter values. These algorithms readily handle mixed aggregations including water heaters, refrigerators, pumps, fans, vacuum systems, *etc*. Due to space limitations, however, this section models air conditioners only.

Fig. 4 illustrates the TCL model. The state is the load temperature T (°C). The input signals are the electric power P (kW) used by the TCL, the surrounding temperature θ (°C), and the thermal power $\dot{Q}$ (kW) from exogenous sources such as the sun, electronics, lights, bodies, *etc.* The parameters are the resistance R (°C/kW), capacitance C (kWh/°C), power capacity $\overline{P}$ (kW), and coefficient of performance η (-).

TABLE I
Baseline Simulation Parameters

Parameter	Value or range
Number of loads, L	50
Time step, Δt	1 minute
Time horizon, $K\Delta t$	24 hours
Thermal resistance, R	2–3 kW/°C
Thermal capacitance, C	1.5–2.5 °C/kWh
Coefficient of performance, η	2.5–3.5
Thermostat deadband halfwidth, ε	0.5 °C
Design outdoor temperature, θ^{des}	40 °C
Design indoor temperature, T^{des}	23–26 °C
Design exogenous thermal power, $\dot{Q}^{\text{des}}$	2.25–3.5 kW
Oversize ratio, ρ	1.5–2.5

The continuous-time temperature dynamics are

$$C\dot{T}(t) = \frac{\theta(t) - T(t)}{R} + \dot{Q}(t) - \eta P(t). \tag{11}$$

This form represents air conditioning, but the model accommodates heating by changing the sign of $\eta P(t)$. Assuming a zero-order hold on the input signals, the discrete-time dynamics are

$$\begin{aligned} T(k+1) = aT(k) + (1-a)[\theta(k) \\ + R(\dot{Q}(k) - \eta P(k))], \end{aligned} \tag{12}$$

where $a = \exp(-\Delta t/(RC))$ and Δt (h) is the time step. The load temperature should satisfy

$$\left|T(k) - T^{\text{set}}(k)\right| \leq \varepsilon, \tag{13}$$

where $T^{\text{set}}(k)$ (°C) is a user-specified setpoint and ε (°C) is the halfwidth of a band of acceptable temperatures.

The TCL model fits into the mathematical framework of Sec. III through the definitions $u_\ell(k) = P(k)/\overline{P}$ and

$$y_\ell(k) = \frac{T^{\text{set}}(k) + \varepsilon(k) - T(k)}{2\varepsilon(k)}. \tag{14}$$

With these definitions, $u_\ell(k) = 0$ if the TCL is off ($P(k) = 0$) and $u_\ell(k) = 1$ if it is on ($P(k) = \overline{P}$), as the framework requires. Furthermore, $y_\ell(k) = 0$ at the hot end of the temperature band ($T(k) = T^{\text{set}}(k) + \varepsilon$) and $y_\ell(k) = 1$ at the cold end ($T(k) = T^{\text{set}}(k) - \varepsilon$), so the temperature constraint (13) is equivalent to the framework's output constraint (2). Finally, all else being equal, $y_\ell(k+1)$ will be larger (the load will be colder) if $u_\ell(k) = 1$ than if $u_\ell(k) = 0$, as required.

The simulation inputs are drawn from independent symmetric triangular distributions over the ranges in Table I. After generating its R, C and η, each air conditioner is sized as $\overline{P} = \rho P^{\text{des}}$, where ρ is the oversize ratio and P^{des} solves (11) in steady state under the design conditions T^{des}, θ^{des} and $\dot{Q}^{\text{des}}$ in Table I. This sizing method follows industry standards [36], which recommend oversizing air conditioners by $\rho = 1.4$. In practice, installers often oversize equipment well beyond this guideline. In the field study [37], for example, Djunaedy *et al.* found air conditioners commonly oversized by $\rho > 2$, with some oversized up to $\rho = 4$. In the simulations here, ρ varies

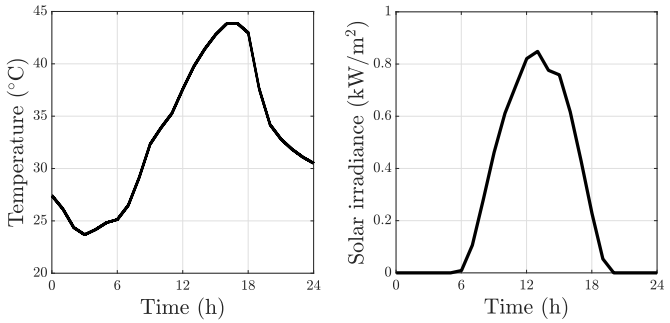

Fig. 5. Outdoor air temperature (left) and global solar irradiance on a horizontal surface (right) on the simulation day. Hour zero is midnight.

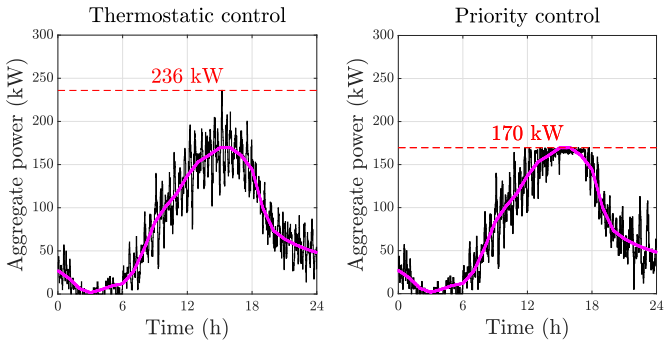

Fig. 6. Aggregate power under thermostatic control (left) and priority control (right). Magenta curves are the aggregate power under perfect variable-speed control. Dashed red lines are demand peaks. Hour zero is midnight.

from 1.5–2.5; the resulting electrical capacities vary from 4–8 kW. Indoor temperature setpoints are drawn from the ranges in Table I. They vary from load to load but are constant over time.

Simulations are set in Sacramento, California, on August 14, 2020, during an extreme heat wave that caused rolling blackouts throughout the state. Fig. 5 shows the outdoor temperature (left) and solar radiation intensity on a horizontal surface (right) on this day. The outdoor temperature peaked at 44 °C (111 °F), well above Sacramento's cooling design temperature of 40 °C. In simulations, all loads are located in the same region, so they see the same outdoor air temperature θ. The thermal power $\dot{Q}$ from the sun and internal heat sources varies from load to load, however, to model diverse occupancy profiles and solar exposures.

Peak shaving is benchmarked here against the variable-speed bound (5). The variable-speed bound is computed by substituting each load's temperature setpoint for the indoor temperature T in (12), solving (12) for each variable-speed load's power consumption, summing the results over loads to compute aggregate demand, and taking the maximum over time to compute the peak. This process emulates variable-speed loads that perfectly track their temperature setpoints.

A. Peak Shaving Performance

Fig. 6 shows the aggregate demand under thermostatic control (left) and priority control (right) over the day. In both plots in Fig. 6, the magenta curve is the aggregate demand under perfect variable-speed control. The dashed red lines in Fig. 6 show the peak demand over the day for each control method. The priority control illustrated in Fig. 6 uses temperature as the priority scoring metric, $s_\ell(k) = -y(k)$, with $y_\ell(k)$ defined by (14). The priority score is set to ∞ if $y(k) < 0$, or if $u_\ell(k-1) = 1$ and the on-time (10) is less than five minutes. This ensures temperature constraint satisfaction and avoids short-cycling. In this simulation, the demand cap P^{cap} is set to the variable-speed bound. Priority control runs continuously, but only differs substantially from conventional bang-bang control near peak hours.

The key message of Fig. 6 is that priority control can reduce peak demand to the variable-speed bound. In this simulation, the variable-speed bound is 170 kW. Under thermostatic control, the peak demand is 236 kW. Priority control reduces the peak demand in this simulation by 28%. This peak reduction is achieved with no loss of temperature regulation performance; the mean absolute temperature error

$$\frac{1}{KL}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\left|T_\ell(k) - T_\ell^{\text{set}}(k)\right|$$

is 0.263 °C under thermostatic control and 0.259 °C under priority control. The energy consumption is nearly identical under thermostatic and priority control. Devices are switched slightly more frequently under priority control (2.82 switches per device per hour) than thermostatic control (2.69).

Taken together, this section's peak reduction result and temperature regulation result suggest that priority control is essentially optimal. Priority control reduces peak demand to the fundamental bound from Sec. IV, and does so while maintaining or improving temperature regulation. This is the largest non-intrusive peak reduction possible; if the peak is reduced further, then not all temperatures can be maintained. To the authors' knowledge, no existing methods have demonstrated optimality in this sense.

This section presents results from a single one-day simulation. However, peak reduction, temperature regulation, energy consumption and device switching were also investigated in the wide range of Monte Carlo simulations described in Sec. VI-C and Sec. VI-D, and in a range of month-long simulations. The key result of this section – that priority control can reduce the peak demand to the variable-speed bound without sacrificing quality of service – was replicated in all 52,000 simulated days.

The peak shaving demonstrated in this section could have significant economic value to customers who pay monthly peak demand charges. For example, according to Table 1 of [1], the highest peak demand price a California utility charged in 2017 was 47.08 $/kW. Under this demand price, the 66 kW peak reduction in this simulation would reduce monthly electricity bills by $3,100, or $62 per load. For comparison, the load-average energy cost over the full month of August, 2020, would have been $125 with an energy price of 0.2 $/kWh.

B. Adaptively Tuning the Demand Cap

The demand cap P^{cap} is an important input to the priority control algorithm. In Sec. VI-A, the demand cap was set *a*

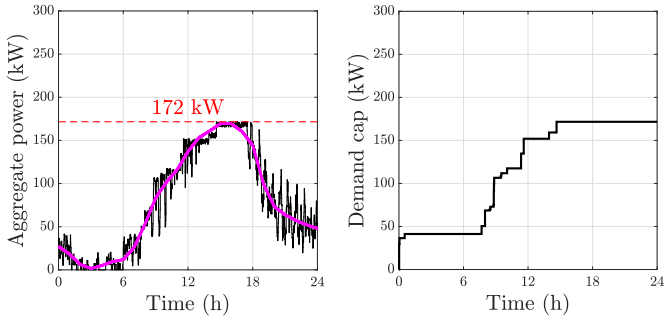

Fig. 7. Aggregate power (left) and demand cap (right) in priority control with the cap initialized at zero and adaptively tuned. The magenta curve is the aggregate power under perfect variable-speed control. Hour zero is midnight.

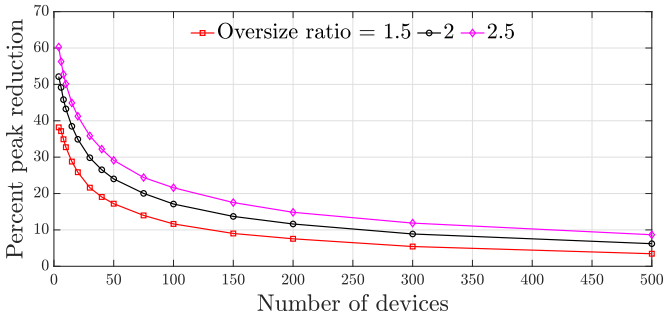

Fig. 8. Scaling of the achievable percent peak reduction (relative to thermostatic control) with the number of devices and oversize ratio.

priori to the variable-speed bound of 170 kW. This method may not be viable in practice, as calculating the variable-speed bound requires exact knowledge of all load models and parameters. A simple but effective alternative is to initialize the cap at some low value, allow it to be violated if necessary to maintain setpoints, and dynamically update the cap if it is violated:

$$P^{\text{cap}}(k+1) = \max\left\{P^{\text{cap}}(k), \sum_{\ell \in \mathcal{L}} P_\ell(k)\right\}.$$

This update rule can be implemented in a fully distributed fashion with no additional communication, as each device can infer all other devices' power states from the information exchanged in the distributed priority control in Sec. (V-C).

Fig. 7 shows the results of priority control with this update rule, initialized with $P^{\text{cap}}(0) = 0$, in the same scenario discussed in Sec. VI-A. The left plot shows that the aggregate power peaks at 172 kW, 1% above the priority control in Sec. VI-A. The right-hand plot shows the evolution of the demand cap, which starts at zero, increases in a series of step changes, and stabilizes after crossing the variable-speed bound of 170 kW. A similar convergence pattern was observed in other simulation runs: the demand cap stabilizes once the system observes a peak event, and the final value is slightly larger than the theoretical limit provided by the variable-speed performance bound. In this simulation, the mean absolute temperature error was 0.261 °C, slightly better than thermostatic control. These results suggest that the demand cap can be adaptively tuned with little impact on performance.

C. Achievable Peak Reduction

Fig. 8 shows the achievable peak reduction as a function of the number of devices L for three fixed values of the oversize ratio ρ. In this figure, the percent peak reduction is defined relative to thermostatic control. Each point in Fig. 8 is the sample-average percent peak reduction over 1,000 Monte Carlo runs. With this sample size, estimates of the mean percent peak reductions are accurate to within $\pm 0.1\%$.

The simulations underlying Fig. 8 use the same input data discussed earlier in this section, except that the oversize ratio is a deterministic parameter. At each (L, ρ) combination, 1,000 one-day Monte Carlo simulations are run. The simulations span 16 values of L, three values of ρ, and two control methods. Fig. 8 therefore summarizes 246 simulated years.

Three conclusions can be drawn from Fig. 8. First, the peak reduction opportunity is significant. In the realistic example of a hotel or apartment building with $L = 50$ devices oversized by a ratio of $\rho = 2$, the peak demand could be reduced by 23%. Second, the opportunity decreases with the number of loads. For example, with $L = 20$ loads and an oversize ratio of $\rho = 2$, the peak could reduced by 31%; while with $L = 500$ loads and the same oversize ratio, the peak could only be reduced by 10%. This can be understood through the Central Limit Theorem. Loosely, the aggregate demand under thermostatic control is the sum of many independent random variables, so its standard deviation (and therefore the probability of a high thermostatic peak) decreases at a rate of $1/\sqrt{L}$. Third, the peak-reduction opportunity increases with the oversize ratio. For example, with $L = 50$ loads oversized by $\rho = 1.5$, the peak could be reduced by 16%; while with $L = 50$ loads oversized by $\rho = 2.5$, the peak could be reduced by 26%. To the authors' knowledge, existing literature has not explored how the achievable peak reduction scales with the number of loads and their degree of oversizing. The understanding of these scaling effects presented here could help practitioners target demand-side management efforts more effectively.

D. On-Time as a Priority Scoring Metric

Table II shows simulation results for four control methods: (1) perfect variable-speed control, (2) thermostatic control, (3) priority control with normalized temperature as the priority scoring metric, and (4) priority control with on-time as the priority scoring metric. Each entry in this table is averaged over 1,000 Monte Carlo runs. In each run, all four control methods are simulated with the same input data. For temperature-based priority control, 16-bit floating-point representation is assumed. The on-time priority controller uses the low-communication implementation discussed in Sec. V-D; each notification is represented using two bits.

Two conclusions can be drawn from Table II. First, on-time is a reasonable priority scoring metric: like temperature-based priority control, on-time-based priority control reduces peak demand to the variable-speed bound with no increase in energy consumption or setpoint tracking error and a mild increase in switching frequency. Second, scoring priority via

TABLE II
AVERAGE RESULTS OVER 1,000 MONTE CARLO SIMULATIONS

Control method	Peak demand (kW)	Total energy (MWh)	Temperature error (°C)	Switches per device per hour	Network-wide data rate (bps)
Thermostatic	218	1.87	0.264	2.63	0
Perfect variable-speed	167 (-23%)	1.87	0	n/a	0
Temperature priority	167 (-23%)	1.87	0.260 (-1.5%)	2.74 (+4.2%)	13.3
On-time priority	167 (-23%)	1.87	0.261 (-1.1%)	2.80 (+6.6%)	0.04 (-99.7%)

on-time, rather than temperature, decreases communication requirements by 99.7%, from a network-wide data rate of 13.3 bits per second to 0.04 bits per second. This suggests that on-time could be an attractive scoring metric if data rates are limited, as may be the case with power-line communication.

VII. CONCLUSION

This paper has developed control methods that enable cyclic loads to cooperatively reduce their peak aggregate demand. In simulations, these methods maintained or improved quality of service relative to today's controllers. This suggests that unlike traditional demand response programs, the framework developed here could be deployed without needing to recruit or pay users. A load could simply be shipped or retrofitted with the control capabilities developed here. Once plugged in, it could discover neighboring loads and begin coordinating with them. The system could work with no central controller.

This paper has focused on non-intrusive peak shaving at the scale of a building, microgrid, or section of a distribution grid. However, cyclic loads could also trade quality of service for deeper, 'intrusive' peak shaving; they could help the grid in other ways, such as price-based load shifting or emergency curtailment; or they could provide services at the scale of a bulk transmission grid. To achieve these ends, the real-time control methods developed here could be combined with a supervisory control system that operates at slower time scales. The framework developed here could also be tested in hardware. Robustness to communication errors and cyberattacks could be investigated in the centralized and distributed architectures proposed here. Other communication architectures, such as one where devices communicate only with nearby neighbors, could also be considered.

APPENDIX A
PROOF OF NON-INTRUSIVE PEAK-SHAVING BOUND

Proving the bound (5) involves a continuous relaxation to the discrete optimization problem (4),

$$\begin{array}{ll} \text{minimize} & \max_{k\in\mathcal{K}} \sum_{\ell\in\mathcal{L}} P_\ell(k) \\ \text{subject to} & \sum_{k\in\mathcal{K}} P_\ell(k) = K\overline{P}_\ell d_\ell,\ \ell\in\mathcal{L} \\ & P_\ell(k)\in[0,\overline{P}_\ell],\ k\in\mathcal{K},\ \ell\in\mathcal{L}, \end{array} \tag{15}$$

with variables $P_\ell(k)$. The only difference between (15) and (4) is that in (15), $P_\ell(k)$ may take on any real value between zero and $\overline{P}_\ell$. This allows the power of each device to vary continuously between zero and its capacity. In (4), by contrast, $P_\ell(k)$ must be either zero or $\overline{P}_\ell$, meaning the device must be either off or on at full capacity. An optimal solution to (15) is

$$\hat{P}_\ell(k) = \overline{P}_\ell d_\ell,\ k\in\mathcal{K},\ \ell\in\mathcal{L}, \tag{16}$$

with optimal peak demand

$$\max_{k\in\mathcal{K}} \sum_{\ell\in\mathcal{L}} \hat{P}_\ell(k) = \sum_{\ell\in\mathcal{L}} \overline{P}_\ell d_\ell. \tag{17}$$

This can be seen by verifying the KKT optimality conditions for an equivalent reformulation of (15):

$$\begin{array}{ll} \text{minimize} & v \\ \text{subject to} & \sum_{\ell\in\mathcal{L}} P_\ell(k) - v \le 0,\ k\in\mathcal{K} \\ & \sum_{k\in\mathcal{K}} P_\ell(k) - K\overline{P}_\ell d_\ell = 0,\ \ell\in\mathcal{L} \\ & -P_\ell(k)\le 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L} \\ & P_\ell(k) - \overline{P}_\ell \le 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L}. \end{array} \tag{18}$$

The KKT conditions for (18) are the following.
Primal feasibility:

$$\begin{aligned} \sum_{\ell\in\mathcal{L}} \hat{P}_\ell(k) - v^\star \le 0,\ k\in\mathcal{K} \\ \sum_{k\in\mathcal{K}} \hat{P}_\ell(k) - K\overline{P}_\ell d_\ell = 0,\ \ell\in\mathcal{L} \\ -\hat{P}_\ell(k)\le 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L} \\ \hat{P}_\ell(k) - \overline{P}_\ell \le 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L}. \end{aligned}$$

Dual feasibility:

$$\begin{aligned} \gamma(k)^\star \ge 0,\ k\in\mathcal{K} \\ \lambda_\ell(k)^\star \ge 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L} \\ \mu_\ell(k)^\star \ge 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L}. \end{aligned}$$

Complementary slackness:

$$\begin{aligned} \gamma(k)^\star \left(\sum_{\ell\in\mathcal{L}} \hat{P}_\ell(k) - v^\star\right) = 0,\ k\in\mathcal{K} \\ \lambda_\ell(k)^\star P_\ell^\star(k) = 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L} \\ \mu_\ell(k)^\star (P_\ell^\star(k) - \overline{P}_\ell) = 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L}. \end{aligned}$$

Stationarity:

$$\begin{aligned} \gamma(k)^\star + \nu_\ell^\star = \lambda_\ell(k)^\star - \mu_\ell(k)^\star,\ k\in\mathcal{K},\ \ell\in\mathcal{L} \\ \sum_{k\in\mathcal{K}} \gamma(k)^\star = 1. \end{aligned}$$

The KKT conditions hold with primal variables (16) and

$$v^\star = \sum_{\ell\in\mathcal{L}} \overline{P}_\ell d_\ell,$$

as well as dual variables

$$\begin{aligned} \gamma(k)^\star = 1/K,\ k\in\mathcal{K} \\ \lambda_\ell(k)^\star = 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L} \\ \mu_\ell(k)^\star = 0,\ k\in\mathcal{K},\ \ell\in\mathcal{L} \\ \nu_\ell^\star = -1/K,\ \ell\in\mathcal{L}. \end{aligned}$$

Thus, (16) is a solution to (15) with optimal value (17).

The optimal value of the fixed-speed problem (4) is denoted by $p^\star$. As (4) and (15) have the same objective function and any solution to (4) is feasible for (15), the optimal value of (15) provides a lower bound on the optimal value of (4):

$$p^\star \geq \sum_{\ell \in \mathcal{L}} \overline{P}_\ell d_\ell. \tag{19}$$

This is the variable-speed bound (5).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of Exelon Corporation and the insight of Theresa Christian, Jake Jurewicz, Sunny Elebua, and the Corporate Strategy team at Exelon. The authors thank Dr. David Blum of Lawrence Berkeley National Laboratory for thoughtful discussions.

REFERENCES

[1] J. McLaren, P. Gagnon, and S. Mullendore, "Identifying potential markets for behind-the-meter battery energy storage: A survey of us demand charges," National Renewable Energy Laboratory, Golden, CO, Tech. Rep. NREL/BR-6A20-68963, 2017.

[2] S. Shabshab, J. Nowocin, P. Lindahl, and S. Leeb, "Microgrid modeling and fuel savings opportunities through direct load control," in *IECON 2018 – 44th Annual Conference of the IEEE Industrial Electronics Society*, 231-236, Ed. IEEE, 2018.

[3] M. Humayun, A. Safdarian, M. Degefa, and M. Lehtonen, "Demand response for operational life extension and efficient capacity utilization of power transformers during contingencies," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 2160–2169, 2014.

[4] J. Eto, "The past, present, and future of us utility demand-side management programs," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-39931, 1996.

[5] S. Iacovella, F. Ruelens, P. Vingerhoets, B. Claessens, and G. Deconinck, "Cluster control of heterogeneous thermostatically controlled loads using tracer devices," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 528–536, 2015.

[6] J. Mathieu, M. Kamgarpour, J. Lygeros, G. Andersson, and D. Callaway, "Arbitraging intraday wholesale energy market prices with aggregations of thermostatic loads," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 763–772, 2015.

[7] F. Ruelens, B. Claessens, S. Vandael, B. De Schutter, R. Babuska, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2149–2159, 2017.

[8] V. Trovato, F. Teng, and G. Strbac, "Role and benefits of flexible thermostatically controlled loads in future low-carbon systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5067–5079, 2018.

[9] I. Jendoubi, K. Sheshyekani, and H. Dagdougui, "Aggregation and optimal management of TCLs for frequency and voltage control of a microgrid," *IEEE Transactions on Power Delivery*, vol. 36, no. 4, pp. 2085–2096, 2020.

[10] D. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy," *Energy Conversion and Management*, vol. 50, no. 5, pp. 1389–1400, 2009.

[11] J. Mathieu, S. Koch, and D. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Transactions on Power Systems*, vol. 28.1, pp. 430–440, 2013.

[12] P. Douglass, R. Garcia-Valle, P. Nyeng, J. Ostergaard, and M. Togeby, "Smart demand for frequency regulation: Experimental results," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1713–1720, 2013.

[13] H. Hao, B. Sanandaji, K. Poolla, and T. Vincent, "Aggregate flexibility of thermostatically controlled loads," *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 189–198, 2015.

[14] V. Lakshmanan, M. Marinelli, J. Hu, and H. Bindner, "Provision of secondary frequency control via demand response activation on thermostatically controlled loads: Solutions and experiences from Denmark," *Applied Energy*, vol. 173, pp. 470–480, 2016.

[15] N. Mahdavi, J. Braslavsky, M. Seron, and S. West, "Model predictive control of distributed air-conditioning loads to compensate fluctuations in solar power," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 3055–3065, 2017.

[16] N. Mahdavi and J. Braslavsky, "Modelling and control of ensembles of variable-speed air conditioning loads for demand response," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4249–4260, 2020.

[17] L. Yao and H. Lu, "A two-way direct control of central air-conditioning load via the Internet," *EEE Transactions on Power Delivery*, vol. 24, no. 1, pp. 240–248, 2008.

[18] S. Lee, S. Kim, and S. Kim, "Demand side management with air conditioner loads based on the queuing system model," *IEEE Transactions on Power Systems*, vol. 26, no. 2, pp. 661–668, 2010.

[19] S. Barker, A. Mishra, D. Irwin, P. Shenoy, and J. Albrecht, "Smartcap: Flattening peak electricity demand in smart homes," in *IEEE International Conference on Pervasive Computing and Communications*. IEEE, 2012, pp. 67–75.

[20] T. Nghiem, M. Behl, R. Mangharam, and G. Pappas, "Green scheduling of control systems for peak demand reduction," in *IEEE Conference on Decision and Control and European Control Conference*. IEEE, 2011, pp. 5131–5136.

[21] G. Karmakar, A. Kabra, and K. Ramamritham, "Coordinated scheduling of thermostatically controlled real-time systems under peak power constraint," in *Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 2013, pp. 33–42.

[22] G. Karmakar and A. Kabra, "Energy-aware thermal comfort-band maintenance scheduling under peak power constraint," in *Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 2013, pp. 122–127.

[23] E. Ancillotti, R. Bruno, and M. Conti, "Smoothing peak demands through aggregate control of background electrical loads," in *Innovative Smart Grid Technologies*, 2014, pp. 1–5.

[24] S. Li, W. Zhang, J. Lian, and K. Kalsi, "Market-based coordination of thermostatically controlled loads – Part I: A mechanism design formulation," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1170–1178, 2016.

[25] J. Wang, A. Raza, T. Hong, A. Sullberg, F. De Leon, and Q. Huang, "Analysis of energy savings of CVR including refrigeration loads in distribution systems," *IEEE Transactions on Power Delivery*, vol. 33, no. 1, pp. 158–168, 2017.

[26] S. Xiang, L. Chang, B. Cao, Y. He, and C. Zhang, "A novel domestic electric water heater control method," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3246–3256, 2020.

[27] M. Franceschelli, A. Pilloni, and A. Gasparri, "Multi-agent coordination of thermostatically controlled loads by smart power sockets for electric demand side management," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 2, pp. 731–743, 2021.

[28] S. Shabshab, P. Lindahl, S. Leeb, and J. Nowocin, "Autonomous demand smoothing for efficiency improvements on military forward operating bases," *IEEE Transactions on Power Delivery*, vol. 35, no. 5, pp. 2243–2251, 2020.

[29] Z. Xu, R. Diao, S. Lu, J. Lian, and Y. Zhang, "Modeling of electric water heaters for demand response: A baseline PDE model," *IEEE Transactions on Smart Grid*, vol. 5, no. 5, pp. 2203–2210, 2014.

[30] K. J. Kircher and K. M. Zhang, "On the lumped capacitance approximation accuracy in RC network building models," *Energy and Buildings*, vol. 104, pp. 454–462, 2015.

[31] K. Kircher, W. Schaefer, and K. Zhang, "A computationally efficient, high-fidelity testbed for building climate control," *Journal of Engineering for Sustainable Buildings and Cities*, vol. 2, no. 1, 2021.

[32] N. Mahdavi, J. Braslavsky, and C. Perfumo, "Mapping the effect of ambient temperature on the power demand of populations of air conditioners," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1540–1550, 2018.

[33] K. Sharma and L. Saini, "Power-line communications for smart grid: Progress, challenges, opportunities and status," *Renewable and Sustainable Energy Reviews*, vol. 67, pp. 704–751, 2017.

[34] A. Aderibole, E. Saathoff, K. Kircher, S. Leeb, and L. Norford, "Power line communication for low-bandwidth control and sensing," *IEEE Transactions on Power Delivery*, 2021.

[35] S. Shabshab, P. Lindahl, J. Nowocin, and S. Leeb, "Voltage waveform transient identification for autonomous load coordination," *IEEE Access*, vol. 7, pp. 123 128–123 137, 2019.

[36] *Manual S: Residential Equipment Selection*, 2nd ed., Air Conditioning Contractors of America, 2015.

[37] E. Djunaedy, K. Van den Wymelenberg, B. Acker, and H. Thimmana, "Oversizing of HVAC system: signatures and penalties," *Energy and Buildings*, vol. 43, no. 2–3, pp. 468–475, 2011.