Chapter 1
Asynchronous coordination of distributed energy resources with packetized energy management

Mads Almassalkhi, Luis Duffaut Espinosa, Paul D. Hines, Jeff Frolik, Sumit Paudyal, and Mahraz Amini

Abstract To enable greater penetration of renewable energy, there is a need to move away from the traditional form of ensuring electric grid reliability through fast-ramping generators and instead consider an active role for flexible and controllable distributed energy resources (DERs), e.g., plug-in electric vehicles (PEVs), thermostatically-controlled loads (TCLs), and energy storage systems (ESSs) at the consumer level. However, in order to facilitate consumer acceptance of this type of load coordination, DERs need to be managed in a way that avoids degrading the consumers quality of service (QoS), autonomy, and privacy. This work leverages a probabilistic packetized approach to energy delivery that draws inspiration from random access, digital communications. Packetized energy management (PEM) is an asynchronous, bottom-up coordination scheme for DERs that both abides by the constraints of the transmission and distribution grids and does not require explicit knowledge of specific DERs local states or schedules. We present a novel macro-model that approximates the aggregate behavior of packetized DERs and is suitable for estimation and control of available flexible DERs to closely track a time-varying regulation signal. PEM is then implemented in a transmission/distribution system setting, validated with realistic numerical simulations, and compared against state-of-the-art load coordination schemes from industry.

Mads Almassalkhi, Luis Duffaut Espinosa, Paul D. Hines, Jeff Frolik, Sumit Paudyal, and Mahraz Amini
University of Vermont, 210 Colchester Avenue, Farrell Hall 200A, Burlington, Vermont 05405 e-mail: {malmassa,lduffaut,phines,jfrolik,mamini2}@uvm.edu

Sumit Paudyal
Department of Electrical and Computer Engineering, Michigan Technological University, 1400 Townsend Drive, Houghton, Michigan 49931 e-mail: sumitp@mtu.edu

1.1 What is Packetized Energy Management (PEM)?

PEM leverages the packet-based strategies from random access communication channels, which have previously been applied to the distributed management of wireless sensor networks [1]. In particular, PEM may be thought of as a multi-channel, multi-receiver version of ALOHA or RTS/CTS (*request/clear to send*) [2], [3]. Under PEM, the delivery of energy to a flexible load (e.g., electric vehicle, water heater, battery, air-conditioner, refrigerator, etc) is accomplished by having the load stochastically request "energy packets," just as digital communication networks break data into packets. An energy packet represents a fixed-duration/fixed-power block of demand consumed (or delivered) by the flexible load. For example, a five minute/ five kW energy packet consumes five kW for five minutes (i.e., 0.417 kWh of energy). PEM engenders the following technical advantages:

- **Local decision-making**: devices offer their own flexibility to the grid operator in bottom-up fashion based on unique local energy demands, which ensures customers' quality of service (QoS).
- **Privacy**: individual energy usage information is not needed, which ameliorates privacy concerns.
- **Fairness**: all devices have equitable access to the grid resources
- **Responsiveness**: the aggregation of devices can adapt to rapid changes in supply and demand.
- **Scalability**: asynchronous control enables plug-and-play capability and scales to millions of devices.

Furthermore, with the proposed PEM coordination architecture, the grid operator (or load aggregator) only requires two scalar measurements from the collection of loads: aggregate power consumption and the loads' requests for packets. This represents a significant advantage over many other load coordination methods that can require up to an entire histogram of states from the population of loads. This distinction is further elaborated upon in Section 1.2.

This chapter will (1) present the PEM scheme within the context of existing approaches from literature, (2) develop and validate an aggregate and homogeneous macro-model, and (3) illustrate coordination of heterogeneous DERs at the consumer level under PEM. PEM is suitable for a large class of deferrable loads and is illustrated with residential electric water heaters, vehicles, and batteries, with an emphasis of content on electric water heaters. Since this chapter focuses on the close-loop control performance of PEM, the underlying communication network is assumed ideal (i.e., no delays of or lost requests or responses). This assumption does not detract from the results presented herein since realistic communication delays and losses are related to an individual device, which is coupled to the system in an asynchronous and randomized manner. Specifically, communication delays are on the order of seconds while packet durations are on the order of minutes. Of course, widespread disruption to the communication infrastructure will affect PEM, which is why implementation of cyber-security and validation against realistic communication parameters are critical topics to consider, but are outside the scope of this

chapter.

1.2 The need for PEM coordination of DERs

Fast-ramping generators have provided the electric grid reliable operating reserves for decades. However, power systems with high penetrations of renewable energy challenge this operating paradigm. At high levels of renewable penetration, current approaches to manage the variability in wind or solar generation would require having more fast-ramping conventional generators online. However, that leads to more generators idling, burning fuel, and increasing harmful air-emissions. Therefore, there is a need to move away from the traditional form of ensuring reliability to consider an active role for flexible and controllable net-load distributed energy resources (DERs), e.g., plug-in electric vehicles (PEVs), thermostatically-controlled loads (TCLs), energy storage systems (ESS's), and distributed generation at the consumer level [4]. While the core concepts underlying modern demand-side management (DSM) have existed for decades [5], [6], the technology for coordinating the activities of DERs is nascent but maturing rapidly. Indeed, there is a growing consensus that balancing supply and demand in power systems with large amounts of variable renewable energy will require an active role for flexible DERs in addition to balancing services from conventional power plants [4].

With the proposed PEM architecture, the grid operator or aggregator only requires two scalar measurements from the collection of loads: the aggregate power consumption and the aggregate request rate. This limited data requirement represents a significant advantage over other aggregate model-estimator-controller state-space approaches, e.g., [7], which requires an entire histogram of states from the collection of loads to update a state bin transition model. To generate these states for control, an observer is designed to estimate the histogram based on aggregated power consumption; however, in some cases, the model may not be observable [8]. Note that in addition to aggregated power consumption, which only informs the observer about devices that are ON, PEM also receives packet requests from the loads that are OFF, which supplies information about the OFF-population and offers a valuable mechanism for observability and state feedback.

Recently, the work in [7] has been extended to include higher order dynamic models and end-user and compressor delay constraints [9] and stochastic dynamical performance bounds [10]. Specifically, the modeling of packet duration in this chapter was inspired by the compressor lock-out method utilized in [9]. A mean-field approach to direct load control is developed for heterogenous TCL populations in [11]. Similarly to the PEM paradigm, the mean-field approach developed in [8], [12] maintains the quality of service (QoS) through opt-out mechanisms and also employs local randomization, which reduces the effect of synchronization in the population of loads. That is, prior work uses either direct load control [4] or employs local randomization at the device-level for the ON/OFF transitions in a

population of flexible loads [7], [12] In the latter case, stochastic device behavior is regulated by either broadcasting (i.e., sending in top-down fashion to the entire population of loads) an updated probability density function [7] or broadcasting an updated scalar variable, which perturbs a probability density functions defined at each device [12]. In contrast to those prior works, PEM does not perturb the probability density function at the device-level nor broadcasts the control signal to the entire population. Instead, PEM listens to each load's individual and stochastic request (i.e., in a locally-driven, bottom-up fashion). The coordinator then responds in real-time to each packet request based on grid and/or market conditions.

The most closely related work on energy packets is found in references [13], [14], where an omniscient centralized *packetized direct load controller* (PDLC) is developed for TCLs. The average controller performance and consumer QoS is analytically investigated and queuing theory is employed by the authors to quantify the centralized controller's performance. In [15], a distributed (binary information) version of PDLC is proposed that requires only (binary) packet request information from the loads. Unlike the proposed PEM scheme, the distributed PDLC assumes that the exact number of participating packetized loads at any given time is known, the allocation of packet requests from the queue is synchronized, and the queue stores packet requests if the packets cannot be allocated, which creates delays in service. Instead, this chapter extends the authors' previous packetized energy results for managing PEVs [16], [17] to consider TCLs and bidirectional residential batteries. In addition, it is shown how PEM uses local randomization and packetization to overcome the challenges surrounding synchronization of devices during extended peak reduction events. Furthermore, PEM does not require the storing of packets requests. More precisely, this chapter focuses on results on PEM coordination of TCLs (specifically, electric water heaters) from [18] and [19] and presents a macro-model of TCLs under PEM. PEM coordination is then extended with a case-study that considers *diverse* heterogeneous loads (TCLs, PEVs, and ESSs). The chapter concludes with a discussion of future directions for PEM in transmission and distribution system operations.

1.3 PEM fundamentals

Figure 1.1 illustrates an example of the cyber-physical interactions needed to realize PEM in distribution system operations, such as managing power constraints, peak demand, and variability from and balancing of renewable generation. We will separately describe the functions of the grid operator (e.g., a utility or ISO), the coordinator (e.g., DER management system or a "virtual power plant" or VPP), and the packetized energy controller (PEC). The PEC connects a single flexible load to a VPP and can directly interact with the load to engender the "packetized" response. Owing to the proposed bottom-up approach, the concept of a packetized load is introduced next.

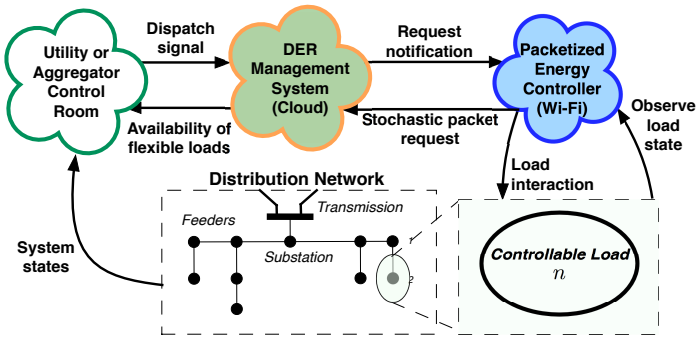

Fig. 1.1: Cyber-physical infrastructure to realize PEM.

1.3.1 Packetized loads

As noted, PEM has previously been proposed for coordinated charging of electric vehicles subject to transformer constraint in a distribution feeder [16], [20], [21]. In this earlier work, PEVs asynchronously request the authority to charge with a specific probability according to their logic state in a probabilistic automaton. For example, consider a three-state finite-state machine (e.g., see Fig. 1.2(right)). The probability that a packetized load in logic state i requests access to the grid during period Δt is P_i where $P_1 > P_2 > P_3$. If there is capacity available in the grid, a PEV's request for a (charging) packet is accepted, and the PEV is granted authority to charge, but only for a pre-determined fixed duration of time (e.g., 15 mins), referred to as the *control epoch*. Upon having the packet request accepted, a logic state transition takes place: $P_i \rightarrow P_{i-1}$, which reduces the mean time-to-request (MTTR). In contrast, if the PEV is denied authority to charge due to insufficient capacity (or overload), the MTTR increases with transition $P_i \rightarrow P_{i+1}$. In this chapter, the PEM concept is adapted for the purpose of managing a set of diverse DER types: TCLs, PEVs, and ESS's by specifically coupling a device's local dynamic state (e.g., temperature or state-of-charge) to the device's MTTR (i.e., the stochastic request rate). With the inclusion of packetized ESS devices, bidirectional power exchanges are considered by distinguishing between charge (consume) and discharge (inject) requests. The diverse DER types are combined under a single VPP and the closed-loop performance is presented in Sec. 1.5.

1.3.1.1 Dynamic modeling of DERs

We will summarize the three DERs models in this section. After developing the necessary models, the stochastic request rate of packetized loads is defined as a function of the dynamic state.

PEV and ESS models:

The dynamic models for PEV and ESS are nearly identical, except that a PEV is inherently mobile and during time periods when they are away from home (or a charging station), the PEV battery's state-of-charge decreases at a rate corresponding to the driving pattern. In addition, unlike an ESS, it is assumed that a PEV cannot inject power to the grid and, therefore, represents a uni-directional energy storage device. A general discrete-time model with sampling time Δt of PEV or ESS battery n's state-of-charge (SOC [kWh], x_n) is summarized by the following:

$$x_n[k+1] = x_n[k] + \Delta t \left(-\eta_{\text{sl},n} x_n[k] + \eta_{\text{ch},n} u_{\text{dis},n}[k] - \eta_{\text{dis,n}} u_{\text{dis},n}[k]\right), \tag{1.1}$$

where $\eta_{\text{sl},n}, \eta_{\text{ch/dis},n}$ represents parameters associated with standing losses and charging/discharging efficiency, respectively. If standing losses are not considered, $\eta_{\text{sl},n} \equiv 0$. The control inputs are charge and discharge rates [kW], which are each bounded. For a PEV, $u_{\text{dis,n}}$ is uncontrollable and reflect the vehicle's away-from-home driving pattern. The SOC is also bounded by battery capacity bounds: $x_n \in \{\underline{x}_n, \bar{x}_n\}$. The ESS is assumed to be subject to an uncontrollable background net-demand process (charging and/or discharging) but that simultaneous charging and discharging is feasible.

TCL Model for Electric Water Heater:

Generic models of TCLs can be found in [4], [7], however, this chapter presents a simple model of an electric water heater that uses a stochastic residential hot-water withdrawal rate to describe the temperature dynamics. Therefore, this section focuses on the electric water heater (EWH) model, which is modeled as a first-order single-heating-element thermodynamic model motivated by [22], [23] but modified to consider a uniform thermal mass and hot-water withdrawal by the consumer in liters/min rather than as a fixed energy loss. Detailed discussions on parameters and hot-water withdrawal rates and events can be found in [18] and [19], but are summarized below.

The temperature at time-step $k+1$ is given by:

$$T_n[k+1] = T_n[k] + \Delta t \left(\frac{\eta_n P_n^{\text{rate}}}{c\rho L} z_n[k] - \frac{(T_n[k] - T_{\text{amb}}[k])}{\tau_n} - \frac{(T_n[k] - T_{\text{inlet}}[k])}{60L} w_n[k] \right) \tag{1.2}$$

where $c = 4.186$ [kJ/kg-°C] and $\rho = 0.990$ [kg/liters] represent specific heat capacity and density of water close to 50°C[1]. L [liters] represents the total capacity of the EWH. Note that $P_n^{\text{rate}}, \eta_n$, and z_n are the heating element power transfer rate [kW], heat transfer efficiency, and the binary ON/OFF logic state ($\Rightarrow z_n \equiv 1/0$) of

[1] Physically, c and ρ vary with water temperature, but this relationship is ignored herein as it does not affect the results or conclusion of PEM's local decision making.

EWH n, respectively[2]. The terms $T_{\text{amb}}, \tau_n, T_{\text{inlet}}$ and w_n are the ambient temperature [°C], time-constant due to ambient insulation losses [s], inlet temperature [°C], and hot water withdrawal rate by consumer n in [liters/min], respectively. The hot water withdrawal rate, w_n, represents the uncontrollable background demand for hot-water and is modeled as a Poisson Rectangular Pulse (PRP) random process as discussed briefly in Section 1.4.1 and [18], [19]. To ensure numerical stability, all simulations use $\Delta t \leq 60$s and are presented in Section 1.5.

1.3.1.2 Conventional control of DERs

The vast majority of existing traditional DERs operate in a binary (ON/OFF) manner and are already controlled by simple state machines. For example, a PEV (when charging) will charge continuously at maximum rating until SOC reaches upper limit and then switch to OFF, while a TCL will change logic state based on temperature deadbands. Specifically, a TCL is controlled to maintain a desired temperature set-point, T^n_{set}, within a temperature dead-band, $T^n_{\text{set}} \pm T^n_{\text{DB}}/2$. The local discrete-time control logic can then be described by the following for TCLs:

$$z_n[k] = \begin{cases} 1, & T_n[k] \leq T^n_{\text{set}} - T^n_{\text{DB}}/2 \\ 0, & T_n[k] \geq T^n_{\text{set}} + T^n_{\text{DB}}/2 \\ z_n[k-1], & \text{otherwise} \end{cases}, \tag{1.3}$$

and for PEVs:

$$z_n[k] = \begin{cases} 1, & x_n[k] < \bar{x}_n \\ 0, & x_n[k] \geq \bar{x}_n \\ z_n[k-1], & \text{otherwise.} \end{cases} \tag{1.4}$$

The battery control logic for an ESS device can be described by similar local logic depending on the operating mode (e.g., peak reduction or arbitrage) but is omitted herein. Thus, the proposed PEM scheme requires only the replacement of the existing state machine with a more sophisticated one (i.e., the equivalent of a firmware upgrade) that interacts with a coordinator/aggregator.

1.3.1.3 Adaptation of PEM for DERs

As discussed previously, the key to enable PEM is the local decision-making of the packetized energy controller (PEC), which observes the physical load's local dynamic state. This state is the temperature for a TCL and the state-of-charge (SOC) for PEV and ESS devices. By coupling a device's dynamic state to a stochastic request rate for accessing the grid, PEM effectively perturbs the ON/OFF transition rate of the device, which, in the aggregate, begets flexibility for the VPP operator.

[2] The binary z_n implies that (1.2) is a hybrid dynamic model.

The description below described the PEM adaptation for an electric water heater (i.e., a TCL), but is straightforward to extend to the other DER types.

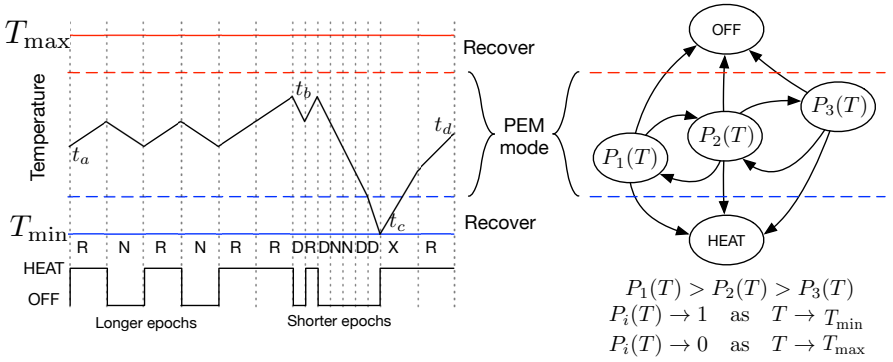

Fig. 1.2: Water heater managed by PEM. The left figure shows a sequence of events. At time t_a, when grid resources are unconstrainted, loads stochastically request (R) or do not request (N) energy. At t_b, the system approaches a period of constrained supply, in which the system aggregator mostly denies requests (D) and reduces the epoch length. As a result, the automaton transitions to a lower probability state (e.g., $P_1 \to P_2$). At t_c, the temperature hits the QoS bound and the load exits (X) from PEM and rapidly seeks to recover temperature to within the QoS bounds, which occurs at t_d. The right figure shows the state machine that changes its request probabilities ($P_i(T)$) and its epoch lengths, based on responses the local temperature state. Also embedded in the automaton is the epoch lengths between state transitions/making requests.

Fig. 1.2(right) illustrates a TCL automaton under PEM. When the local temperature of the TCL, T, is between its upper and lower temperature limits for PEM operation, the TCL's mean time-to-request (MTTR) is driven by an exponential distribution whose mean is inversely proportion to T relative to the upper limit. That is, TCLs with temperatures very close to the lower threshold will make requests with near certainty (i.e., $P_i(T \to T_{\min}) \approx 1$) and those near the upper limit in temperature will make requests with low probability (i.e., $P_i(T \to T_{\max}) \approx 0$). Upon transmitting a request and, if there is capacity in the grid, the TCL will be given the authority to turn ON for a fixed control epoch length δ (i.e., $z_n(t) = 1$ for $t \in (t_0, t_0 + \delta)$), and a state transition occurs: $P_i(T) \to P_{i-1}(T)$. If the request is denied, the TCL finite state machine transitions to a state with lower MTTR, $P_i(T) \to P_{i+1}(T)$, but will immediately resume requesting with temperature-dependent probability. If access is denied repeatedly, T reaches $T_{\min}$, which causes the TCL to exit the PEM scheme (*exit-ON*) to satisfy quality of service (QoS) constraints. An illustrative ON/OFF cycle of a packetized water heater is illustrated in Fig. 1.2(left).

In addition to the TCL receiving an "Yes/No" response to a request, the TCL may also receive an updated (global) control epoch length, δ, thus enabling tighter tracking in the aggregate, which is helpful during ramping events. Clearly, while the TCL is ON, it does not make requests. Furthermore, we require $\delta \geq \Delta t$.

Remark 1.1. Since all packetized loads operate in this manner, the DER aggregator granting ("Yes") or denying ("No") the authority to turn ON does not require any

knowledge of a particular load. Furthermore, the aggregator does not even need to track which load is making a particular request. As each load type runs the same automaton logic and its ability to turn ON depends only on the (present and past) system capacity (and, potentially, past VPP decisions), any load making a request at the same point in time will be treated the same by the aggregator. As such, the PEM approach inherently maintains privacy while still offering equitable access to the grid.

1.3.1.4 The stochastic request rate with PEM

This section explicitly defines the stochastic request rate for a packetized load as a function of the device's local dynamic state. Consider a TCL, PEV, or ESS packetized load with just one automaton state. Then, in the discrete-time implementation of PEM, the probability that the packetized load n with local dynamic state $x_n[k] \in [\underline{x}_n, \overline{x}_n]$ and desired set-point $x_n^{\text{set}} \in (\underline{x}_n, \overline{x}_n)$ requests access to the grid during time-step k (over interval Δt) is defined by the cumulative exponential distribution function:

$$P(x_n[k]) := 1 - e^{-\mu(x_n[k])\Delta t}$$

where rate parameter $\mu(x_n[k]) > 0$ is dependent on the local dynamic state. For a consume (or charge) request, this dependence is established by considering the following boundary conditions:

- $P(\text{load } n \text{ requests to consume packet during time } k \,|\, x_n[k] \leq \underline{x}_n) = 1$
- $P(\text{load } n \text{ requests to consume packet during time } k \,|\, x_n[k] \geq \overline{x}_n) = 0,$

which permits a simple functional form for the rate parameter that ensures the boundary conditions are met:

$$\mu(x_n[k]) = \begin{cases} 0, & \text{if } x_n[k] \geq \overline{x}_n \\ m_R \left(\frac{\overline{x}_n - x_n[k]}{x_n[k] - \underline{x}_n} \right) \cdot \left(\frac{x_n^{\text{set}} - \underline{x}_n}{\overline{x}_n - x_n^{\text{set}}} \right), & \text{if } x_n[k] \in (\underline{x}_n, \overline{x}_n) \\ \infty, & \text{if } x_n[k] \leq \underline{x}_n \end{cases} \tag{1.5}$$

where $m_R > 0$ [Hz] is a design parameter that defines the mean time-to-request (MTTR). For example, if one desires a MTTR of 5 minutes when $x_n[k] \equiv x_n^{\text{set}}$ then $m_R = \frac{1}{600}$Hz.

Figure 1.3 illustrates the bidirectional stochastic request rates and MTTR for a generic packetized load that can request to consume power from and inject power into the grid. Note that (1.5) is represented by the blue lines (left to right).

1.3.1.5 Quality of service under PEM

With the stochastic nature of DERs under PEM, it is entirely possible that a local disturbance (e.g., a large hot water withdrawal rate for TCLs) can drive $x_n[k]$ below

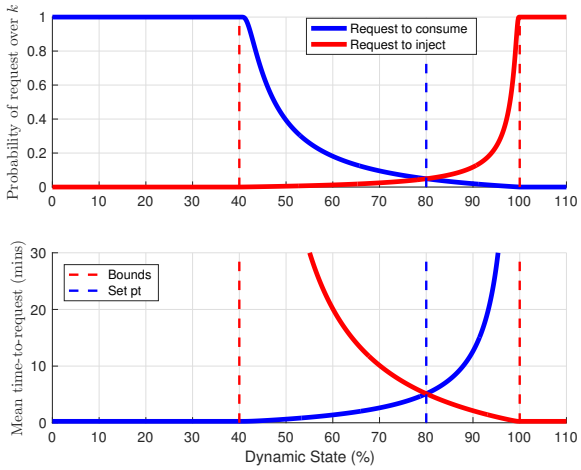

Fig. 1.3: Illustrating the effect of the local state x_n (e.g., temperature or state-of-charge) on the access request probabilities (top) and MTTR (bottom) of a controllable load under under PEM. Note that if the state exceeds bounds $(\underline{x}_n, \overline{x}_n)$, the probability of request is 0 or 1 depending on the type of request (consume or inject). While TCLs can only consume power from the grid, controllable ESS batteries can discharge and inject power into the grid. The ability of a device to request either form of participation is captured with the consume and inject packets. We assume that if both packets are requested simultaneously that the requests cancel eachother out and no request is made.

$\underline{x}_n$. Therefore, to maximize QoS to the consumer (e.g., avoid cold showers), a DER under PEM can temporarily exit (i.e., opt-out of) PEM and operate under traditional control (e.g., turn ON and stay ON). This is illustrated for TCLs in Fig. 1.2(left) at event t_c and with automaton states HEAT and OFF in Fig. 1.2(right). That is, once a TCL under PEM exceeds temperature bounds, the traditional control logic is employed temporarily to bring the local temperature within PEM "recovery bounds" $T^n_{\text{set}} \pm T^n_{\text{PEM}}/2$ with $T^n_{\text{PEM}} < T^n_{\text{DB}}$ where PEM control logic is reinstated (i.e., TCL opts back into PEM). The recovery bounds are helpful to avoid excessive exit/re-entry cycling at the boundary.

Remark 1.2. Clearly, if packetized loads exit PEM *en-masse*, the available flexibility can be greatly reduced and, therefore, will impact ability of a coordinator to track a given reference balancing signal. Hence, the macro-model effort in Section 1.4 is focused on developing a population model of homogeneous TCLs that will permit analysis to leverage incoming requests (modulating yes/no response rates) to reduce the need for opting out without sacrificing controllability.

1.3.2 Closing the loop on PEM with the virtual power plant

As shown in Fig. 1.1, a packetized energy controller enables bidirectional Wi-Fi communication between a load and the virtual power plant (VPP). The VPP receives balancing dispatch signals akin to Automatic Generation Control (AGC) from a grid operator and coordinates flexible energy resources to track the balancing command[3]. Within the proposed PEM scheme, the VPP tracks the balancing signal by responding to individual, asynchronous, and stochastic load access requests with "Yes" or "No" notifications based on real-time output error between actual aggregate output, $y(t)$, and the VPP reference signal, $r(t)$: $e(t) := r(t) - y(t)$. This simple closed-loop controller is illustrated in Fig. 1.4. The VPP is similar to a relay controller in the sense that it accepts a request ("Yes") if $e(t) > 0$, otherwise, "No." However, unlike standard relay control of a single plant, the VPP responds to asynchronous, stochastic requests from N plants, which overcomes common drawbacks associated with relay control (e.g., switching leading to oscillations) and permits accurate tracking. While the above describes a simple control scheme for VPP, more advanced approaches can leverage past load requests rates, VPP responses, and aggregate net-demand of the VPP to further improve upon performance. The VPP is described by the following inputs and outputs:

Input: Balancing reference signal, asynchronous requests;
Output: Yes/No access notification to individual load.

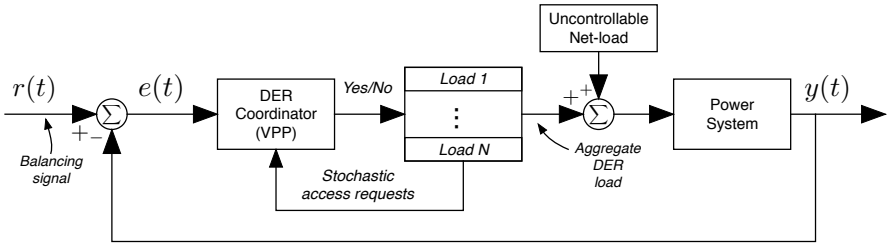

Fig. 1.4: The closed-loop feedback system for PEM with the reference $r(t)$ provided by the Grid Operator and the aggregate packetized loads' output response $y(t)$ measured by VPP.

1.3.3 Providing grid-level service with PEM

The transmission (e.g., ISO New England) or distribution utility system operator (e.g., the DSO Control Room in Fig. 1.1) is able to measure or estimate the states of the grid, such as voltage, frequency, and power flows. Under scenarios with high

[3] While the VPP needs to estimate and predict the aggregate flexibility from available loads, these results focus on the tracking control problem as the estimation problem represents ongoing research

penetration of renewable energy, the grid operator will find it ever more difficult to balance demand and supply while satisfying network conditions and, therefore, seeks to leverage the flexible packetized DERs sitting in customer homes and industrial/commercial facilities. This is achieved by signaling individual balancing requests to VPPs across the network in near real-time akin to Automatic Generation Control (AGC; secondary frequency control) signals, which are transmitted every 4-5 seconds today. The signaling may be computed via solving an optimal power flow (OPF) problem that dispatches VPPs optimally with respect to network constraints and available net-load resources. Thus, the grid operator is summarized by the following inputs and outputs:

Input: Grid voltages, frequencies, net-load forecasts, price forecasts;
Output: Dispatch balancing signal.

Remark 1.3. By managing the anonymous, fair, and asynchronous pings of packetized loads via a VPP that receives grid or market-based balancing signals from the grid operator, PEM represents a bottom-up distributed control scheme that has been adapted for TCLs, PEVs, and ESS devices in this section.

The next section develops and validates a macro-model that captures the aggregate behavior of a population of homogeneous EWHs. This is particularly valuable since the complexity of a large-scale VPP increases exponentially when augmenting the hybrid dynamics in (1.1) and (1.2) for thousands of flexible loads under PEM. Devising suitable control techniques, therefore, becomes intractable for the proposed VPP and a simpler, scalable, lower-dimensional, aggregate model is needed.

1.4 Macro-model for homogeneous packetized EWHs

This section presents a state bin transition (macro)model for a large homogeneous population of TCLs. The aggregate energy use of these TCLs is coordinated with PEM. First, a macro-model for a population of TCLs is developed and then augmented with a timer to capture the duration and consumption of energy packets and with exit-ON/OFF dynamics to ensure consumer quality of service. This permits a virtual power plant (VPP) operator to interact with TCL population through the stochastic packet request mechanism. The VPP regulates the proportion of accepted packet requests to allow tight tracking of balancing signals. The developed macro-model compares well with (agent-based) micro-simulations of TCLs under PEM and can be represented by a controlled Markov chain. Details on the macro-model development can be found in [19].

1.4.1 Conventional thermostatic control

A macro-model for a large population of TCLs is developed in this section as an abstraction of the augmented (agent-based) dynamic micro-models. Specifically, consider the TCL population dynamics over a discretization of the temperature state space and employ a state bin transition model, such that the macro-model approaches the behavior of the micro-model as the number of devices increases [24]. The transitions between these bins are determined by the dynamical system equations of the homogeneous TCLs as discussed below. The macro-model utilizes a finite set $\mathscr{X} = \{x_1, \ldots, x_N\}$, where each element is called a state. Assume that there exists an appropriate probability space $(\Omega, P, \mathscr{F})$, where Ω is the set of events, $\mathscr{F}$ a filtration, and P the probability measure of elements in $\mathscr{F}$. Then, random variables $\{X_k\}_{k\geq 0}$ are defined as $X_k : \Omega \to \mathscr{X}$. Let $x_j \in \mathscr{X}$ and denote $q_j[k] = P(X_k = x_j)$ as the probability of $X_k = x_j$, $k \geq 0$. The column vector $q[k] := (q_1, \ldots, q_N)^T$ then gives the probability mass function of the random variable X_k. Also, if one denotes the transition probability of an homogeneous Markov chain as $p_{ij} = P(X_{k+1} = x_i | X_k = x_j)$, it then follows that

$$q[k+1] = Mq[k], \tag{1.6}$$

where $M = \{p_{ij}\}_{1\leq i,j\leq N}$ [25]. Given an initial distribution $q[0]$, one can solve for (1.6) and find the distribution at time k as $q[k] = M^k q[0]$.

Conventional thermostatic control of an EWH is based on keeping the local state variable (e.g., temperature) within a deadband $[T_{\min}, T_{\max}]$ by switching the device ON (when $T \leq T_{\min}$) or OFF (when $T \geq T_{\max}$), where $[T_{\min}, T_{\max}] = [T_{\text{set}} - T_{\text{DB}}, T_{\text{set}} + T_{\text{DB}}]$ as in Section 1.3. The interval $[T_{\min}, T_{\max}]$ is divided into N consecutive bins each corresponding to a *bin state* in $\mathscr{X}$. Since (1.2) includes hybrid ON/OFF dynamics, the state space for the system consists of two discrete state spaces: $\mathscr{X}_{\text{on}}$ and $\mathscr{X}_{\text{off}}$. That is, the full state space is given by $\mathscr{X} = \mathscr{X}_{\text{on}} \cup \mathscr{X}_{\text{off}}$. At time k, the probability mass function of the system is $q^\top = (q_{\text{on}}^\top, q_{\text{off}}^\top)$ with $q_{\text{on}} = (q_{\text{on}}^1, \cdots, q_{\text{on}}^N)^\top$ and $q_{\text{off}} = (q_{\text{off}}^1, \cdots, q_{\text{off}}^N)^\top$. Note that q contains the percentage of the population in each state of $\mathscr{X}$. For instance, if R is the total number of EWHs and R_{on}^i is the number in state x_{on}^i, then $R_{\text{on}}^i = q_{\text{on}}^i R$. Similarly, the total ON-population is given by

$$y = c^\top q \;\text{ for }\; c = (\mathbf{1}_N^\top, 0 \cdots 0)^\top \in \mathbb{R}^{2N}, \tag{1.7}$$

and $\mathbf{1}_N = (1, \ldots, 1)^\top \in \mathbb{R}^N$. The transition rates are computed by considering how the temperature bin corresponding to a particular state is altered by the hybrid dynamics in (1.2).

Together with discrete sampling time and temperature bin widths, the hot water withdrawal rate w_n in (1.2) is one of the main factors affecting these transition rates. For this purpose, a *Poisson Rectangular Pulse* stochastic differential model is employed to model the individual water withdrawal rates w, including the stochastic

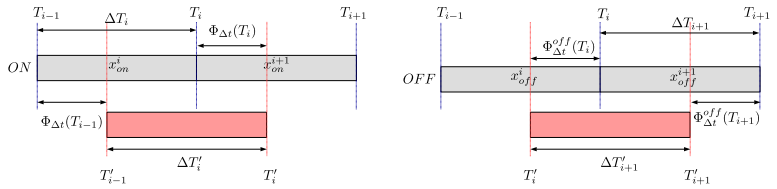

Fig. 1.5: Transition rates calculation for ON and OFF populations.

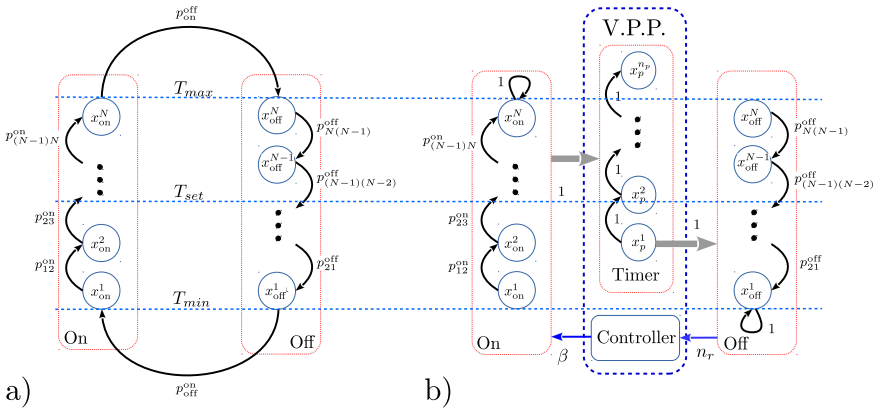

Fig. 1.6: Abstraction for (a) conventional thermostatic control and (b) PEM control, where self-loops are not visualized.

duration and intensities (e.g., l/min) of hot-water events [26]. Upon aggregating w across the entire population, it is shown in [19] that a constant steady-state water consumption rate can be derived, $\bar{w}_{\text{sst}}$. For example, consider two realizations a, b of the water profiles with identical parameters except for the water withdrawal intensities of the random variable w ($\lambda_a \neq \lambda_b$). An EWH in ON state with λ_a ($< \lambda_b$) at temperature T_i reaches temperature T_{i+1} faster than realization b, which draws more hot water on average and increases the time required to reach T_{i+1}. Nevertheless, since the hot water draw profiles in the population are assumed to be statistically identical, the average of these profiles reaches $\bar{w}_{\text{sst}}$ for $t \to \infty$. Thus, the state transition rates for the large population are calculated considering the evolution of (1.2) with respect to the average hot water draw of the population. The transition rates for the ON and OFF populations are computed next. Dropping the subscript n in (1.2), it follows that the solution with steady state consumption $w = \bar{w}_{\text{sst}}$ and $T(0) = T_0$ is

$$T(t) = \Phi_{T_0}(t) = e^{-at}\left(T_0 - \frac{b(z)}{a}\right) + \frac{b(z)}{a}, \tag{1.8}$$

where $a = \frac{1}{\tau} + \frac{\bar{w}_{\text{sst}}}{60L}$ and $b(z) = \frac{T_{amb}}{\tau} + \frac{T_{\text{in}}}{60L}\bar{w}_{\text{sst}} + \frac{P^{rate}}{c\rho L\eta}z$. In particular, define $\Phi^{\text{on}}_{T_0}(t) = \Phi_{T_0}(t)\,|_{z=1}$ and $\Phi^{\text{off}}_{T_0}(t) = \Phi_{T_0}(t)\,|_{z=0}$. For the ON population, the dynamics imply forward transitions, i.e., from x^i_{on} to x^{i+1}_{on} as shown in Fig. 1.5. First, take the bound-

aries of the temperature bin T_{i-1} and T_i corresponding to state x^i_{on} and compute $T'_{i-1} = \Phi^{\text{on}}_{\Delta t}(T_{i-1})$ and $T'_i = \Phi^{\text{on}}_{\Delta t}(T_i)$. Note that in this case $T_i < T'_i$. Thus, the percentage of water heaters that remain in x^i_{on} and move to x^{i+1}_{on}, respectively, are given by

$$p^{\text{on}}_{ii} = \frac{T_i - T'_{i-1}}{T'_i - T'_{i-1}} \quad \text{and} \quad p^{\text{on}}_{i(i+1)} = \frac{T'_i - T_i}{T'_i - T'_{i-1}}.$$

Note that $p^{\text{on}}_{ii} + p^{\text{on}}_{i(i+1)} = 1$, as expected. Transition rates for the OFF dynamics are determined similarly, but are reversed, i.e., from x^{i+1}_{off} to x^i_{off} since $T'_i = \Phi^{\text{off}}_{\Delta t}(T_i) < T_i$. Thus,

$$p^{\text{off}}_{ii} = \frac{T'_{i+1} - T_i}{T'_{i+1} - T'_i} \quad \text{and} \quad p^{\text{off}}_{i(i-1)} = \frac{T_i - T'_i}{T'_{i+1} - T'_i}.$$

The previous analysis was purposely restricted to state transitions between contiguous states. Using (1.8), one can compute an upper bound for Δt such that any EWH in state x^i_{on} only transitions to x^{i+1}_{on} and any EWH in x^{i+1}_{off} only transition to x^i_{on} for all i. Define

$$t^{\text{on}}_i = a^{-1} \log \left(\frac{T_i - \frac{b(z)}{a}}{T_{i+1} - \frac{b(z)}{a}} \right) \Bigg|_{z=1} \tag{1.9}$$

as the time that an EWH takes to go from T_i to T_{i+1}. Observe that if an EWH at T_i is kept ON for $t > t^{\text{on}}_i$ seconds, then $T(t) > T_{i+1}$. This implies that some EWHs in x^i_{on} will transition to x^{i+2}_{on} and skip x^{i+1}_{on}. Similarly, t^{off}_i is defined as in (1.9) but $z = 0$ and the transitions are in a reverse direction. The condition on the discretization time step Δt for *contiguous transitions* is then formulated as $\Delta t < \min_i \{t^{\text{on}}_i, t^{\text{off}}_i\}$. For example, a state space partitioning having $N = 30$ temperature bins for each of the ON and OFF populations implies that $\Delta t < 60.27$ seconds in order to keep transitions between contiguous states.

In addition, the OFF-to-ON ($p^{\text{on}}_{\text{off}}$) and ON-to-OFF ($p^{\text{off}}_{\text{on}}$) transition rates must be computed to capture the jump to a transitory state that automatically transitions to x^1_{on} and x^N_{off}, respectively. The complete Markov chain for conventional thermostatic control is shown in Fig. 1.6a. It is important to observe that the transient effects on temperature caused by stochastic hot water withdrawals are not captured since the transition rates assume a steady state (mean) consumption of hot water. The Markov transition matrix M associated to conventional thermostatic control is then given as

$$M = \left(\begin{array}{cccc|ccc|c} p_{11}^{\text{on}} & 0 & \cdots & 0 & p_{\text{on}}^{\text{off}} & \cdots & 0 & 0 \\ p_{12}^{\text{on}} & p_{22}^{\text{on}} & \ddots & 0 & 0 & \cdots & 0 & 0 \\ 0 & p_{23}^{\text{on}} & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{NN}^{\text{on}} & 0 & \cdots & 0 & 0 \\ \hline 0 & 0 & & 0 & p_{11}^{\text{off}} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \ddots & p_{(N-1)(N-2)}^{\text{off}} & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & p_{(N-1)(N-1)}^{\text{off}} & p_{N(N-1)}^{\text{off}} \\ 0 & 0 & \cdots & p_{\text{off}}^{\text{on}} & 0 & \cdots & 0 & p_{NN}^{\text{off}} \end{array}\right). \quad (1.10)$$

Observe that the Markov chain associated to M is irreducible since one can reach any state from any arbitrarily initial state. It follows then that this abstraction possess a unique invariant distribution as well since $\mathscr{X}$ is finite dimensional. Nonetheless the conventional model lacks the flexibility inherent to PEM.

1.4.2 PEM Markov model

Recall from Section 1.3 that, under PEM, an EWH can only switch ON for an epoch if its packet request is accepted by the VPP coordinator. That is, given the aggregate request rate, the VPP selects the proportion of EWHs that will receive a packet and automatically switch them ON. To capture the unique nature of PEM's fixed packet duration and VPP's role, we leverage prior literature on fault tolerant recovery logic [27] and TCL modeling with compressor lockout periods [9]. Specifically, a fixed *timer* is added to the state bin transition model to track the population with accepted packet requests. The objective of this section is to present a macro-model describing PEM control as a *controlled Markov chain*.

Definition 1.1 (Controlled Markov Chain [25]). Let $\{u_k\}_{k\geq 0}$ be a sequence of real valued functions taking values on a set U. A Markov chain $\{X_k\}_{k\geq 0}$ is said to be a *controlled Markov chain* (CMC) if its transition matrix $M = M(u) := \{q_{ij}(u)\}_{1\leq i,j\leq N}$ satisfies

$$\begin{aligned} &P(X_{n+1} = x_{i_{n+1}} | X_n = x_{i_n}, \ldots, X_0 = x_{i_0}, u_n, \cdots, u_0) \\ &\quad = P(X_{n+1} = x_{i_{n+1}} | X_n = x_{i_n}, u_n) = p_{i_{n+1} i_n}(u_n). \end{aligned}$$

The definition also implies that $M(u)$ for a CMC must be a (column) stochastic matrix for any choice of $u \in U$. Assuming that all states of the CMC are observed, one can define a control policy: $u = \mathscr{X} \to U$, and, thus, $M = M(u(x))$. The probability mass function of a CMC is computed similarly using $q[k+1] = M(u[k])q[k]$ given an initial distribution $q[0]$ and control input $u[0]$.

In what follows, PEM control will be introduced in the context of CMCs. The underlying Markov transition matrix over which PEM is implemented is given by (1.10), but with $p_{\text{off}}^{\text{on}} = p_{\text{on}}^{\text{off}} = 0$ and $p_{NN}^{\text{on}} = p_{11}^{\text{off}} = 1$. That is, x_{off}^1 and x_{on}^N are absorbent states indicating that ON states can not be reached from OFF states and vice-versa. VPP control, therefore, becomes the interface between these two modes of operation. The mechanics of switching EWHs ON and OFF under PEM control are described next.

Suppose $q[k] \in \mathbb{R}^{2N}$ is the probability distribution of the PEM macro-model population at time k, $\beta_{\text{on}} = \text{diag}\{\beta_{\text{on}}^1, \ldots, \beta_{\text{on}}^N\}$ with $\beta_{\text{on}}^i \in [0,1]$ the percentage of the OFF-population in state x_{off}^i that is switched ON by the VPP, and $\beta_{\text{off}} = \text{diag}\{\beta_{\text{off}}^1, \ldots, \beta_{\text{off}}^N\}$ with $\beta_{\text{off}}^i \in [0,1]$ the percentage of the ON-population in state x_{on}^i that is switched OFF. The action of instantaneously switching ON and OFF proportion of devices in q is given by the transformation:

$$\bar{q}[k] = \bar{M}(\beta_{\text{on}}[k], \beta_{\text{off}}[k])\, q[k], \tag{1.11}$$

where

$$\bar{M}(\beta_{\text{on}}, \beta_{\text{off}}) = \left(\begin{array}{c|c} I_N - \beta_{\text{off}} & \beta_{\text{on}} \\ \hline \beta_{\text{off}} & I_N - \beta_{\text{on}} \end{array}\right), \tag{1.12}$$

and I_N denotes the N-dimensional identity matrix. Once $\bar{M}(\beta_{\text{on}}, \beta_{\text{off}})$ has switched some EWHs ON and some other OFF, the underlying transition matrix M acts on $\bar{q}$. This provides the dynamics

$$q[k+1] = M\bar{M}(\beta_{\text{on}}, \beta_{\text{off}})\, q[k]. \tag{1.13}$$

The next theorem simply says that the sequence $\{X_k\}_{k\geq 0}$ associated (1.13) is a CMC.

Theorem 1.1. *Let $\beta_{\text{on}}[k], \beta_{\text{off}}[k] \in \mathbb{R}^{N\times N}$ be defined as in (1.11) for all $k \geq 0$. The sequence $\{X_k\}_{k\geq 0}$ of random variables X_k taking values in $\mathscr{X}$ and probability distribution satisfying (1.13) is a controlled Markov chain as described by Definition 1.1 with input $u[k] = (\mathbf{1}_N^\top \beta_{\text{on}}[k], \mathbf{1}_N^\top \beta_{\text{off}}[k])^\top \in \mathbb{R}^{2N}$.*

Proof: The proof is straightforward since (1.10) and (1.12) are stochastic matrices for arbitrary $\beta_{\text{on}}^i, \beta_{\text{off}}^i \in [0,1]$ for all i, and the product of stochastic matrices is a stochastic matrix. ■

An important aspect of PEM control is that only EWHs that are OFF request a packet and do so as a function of the (local temperature) bin, which implies that not all OFF EWHs will turn ON. Therefore, define

$$\hat{q}[k] := \hat{M}q[k] = \left(\begin{array}{c|c} I_N & 0_N \\ \hline 0_N & T_{\text{req}} \end{array}\right) q[k] = \begin{pmatrix} q_{\text{on}}[k] \\ \hat{q}_{\text{off}}[k] \end{pmatrix},$$

where $T_{\text{req}} = \text{diag}\{p_1^{\text{req}}, \ldots, p_N^{\text{req}}\}$ and $p_i^{\text{req}} := 1 - e^{-\mu(T_i^m)\Delta t}$ is the request probability assigned to x_{off}^i by (1.5) with respect to the mid-point of temperature bin i, T_i^m. Note that $\hat{q}$ is not a probability mass function since $\mathbf{1}_N^\top(q_{\text{on}} + \hat{q}_{\text{off}}) < 1$, which means that the aggregate request rate, i.e., the population that can be switched ON, is given by:

$$n_r[k] := \mathbf{1}_N^\top \hat{q}_{\text{off}}[k]. \tag{1.14}$$

Under PEM, the VPP determines the rate of accepting packets, $\beta[k]$. The resulting EWHs instantly switch ON when packet requests are accepted. The population of devices that switch from OFF to ON, q^+, is a function of β and q_{off}. That is,

$$q^+[k] := \left(\begin{array}{c|c} 0_N & \beta[k]T_{\text{req}} \\ \hline 0_N & -\beta[k]T_{\text{req}} \end{array}\right) q[k] = M^+_{\beta[k]}\, q[k] \tag{1.15}$$

In contrast, to model the population of EWHs that switch from ON to OFF requires information on the rate of expiring packets. In other words, let δ [secs] be the duration of a packet epoch, then the EWHs that have been ON for δ seconds will turn OFF. This requires keeping track of how many EWHs were turned ON δ seconds ago and, essentially, constitutes a delayed system. However, one can augment states to the system dynamics to account for the needed memory, which is equivalent to having a timer. That is, given δ, the time step Δt, and the vector of augmented (timer) states $x_p \in \mathbb{R}^{n_p}$ with $n_p = \lfloor \delta/\Delta t \rfloor$, the *timer dynamics* is given by

$$x_p[k+1] = M_p x_p[k] + C_p q^+_{\text{on}}[k] \tag{1.16a}$$

$$y_p[k] = x_p[k], \tag{1.16b}$$

where $M_p \in \mathbb{R}^{n_p \times n_p}$ is a zero matrix except for its first lower diagonal whose components are 1 and $C_p \in \mathbb{R}^{n_p \times N}$ is responsible for allocating the newly switched ON population into the timer states. Note that there exists a temperature T_p such that $\Phi^{\text{on}}_{T_p}(\delta) = T_{\max}$. Therefore it is necessary for C_p to interrupt packets to prevent exceeding temperature limits and thus wasting resources. Specifically, if $T_{i+1} < T_p$, C_p allocates all requesting EWHs from bin $[T_i, T_{i+1}]$ into the timer state x_p^1. Otherwise, it allocates EWHs with $T_j > T_p$ in the timer state x_p^j with $j = \lfloor (\delta - t_j)/\Delta t \rfloor$ and t_j is the time it will take to increase the EWH's temperature from T_j to $T_{\max}$. Note that since the macro-model considers only binned (rather than exact) temperatures, the allocation of requests assumes that the mass function in each state is uniformly distributed.

The timer states are internal states of the VPP and provide information of the distribution of total ON population in PEM (i.e., $1_N^\top q_{\text{on}}$) across all packet intervals, x_p. As in (1.15), one can define the population of EWHs that just completed their δ-second packet and will turn OFF instantly as

$$q^-[k] := \left(\begin{array}{c|c} \beta^-[k]I_N & 0_N \\ \hline -\beta^-[k]I_N & 0_N \end{array}\right) q[k] = M^-_{\beta^-[k]}\, q[k], \tag{1.17}$$

where $\beta^-[k] := y_p^{n_p}[k]/(\mathbf{1}_{n_p}^\top y_p[k])$. One can now formulate the ON/OFF switching events for the entire population as:

$$\bar{q}[k] := q[k] + q^+[k] - q^-[k] = (I + M^+_{\beta[k]} - M^-_{\beta^-[k]})\, q[k],$$

which yields the EWH population dynamics:

$$\begin{aligned} q[k+1] &= M(I + M^+_{\beta[k]} - M^-_{\beta^-[k]})\, q[k] \\ &= \bar{M}(\beta_{\text{on}}[k], \beta_{\text{off}}[k])\, q[k], \end{aligned} \tag{1.18}$$

where $\beta_{\text{on}}[k] = \beta[k]T_{\text{req}}$ and $\beta_{\text{off}}[k] = \beta^-[k]I_N$. Note that there is no order in which EWHs are switched ON or OFF since both happen simultaneously. Fig. 1.6b shows the state diagram of the population model under PEM control.

The next corollary follows directly from Theorem 1.1.

Corollary 1.1. *The sequence $\{X_k\}_{k\geq 0}$ of random variables X_k taking values in $\mathscr{X}$ and probability distribution satisfying (1.18) is a controlled Markov chain with input $u[k] = (\mathbf{1}_N^\top \beta[k]T_{\text{req}}, \mathbf{1}_N^\top \beta^-[k]I_N)^\top$.*

1.4.3 Tracking with PEM macro-model

In PEM, the input β is exogenous. Recall $P^{\text{rate}} := \frac{1}{n}\sum_{n=1}^N P_n^{\text{rate}}$, P_{ref} and P_{dem} (see [18] for a list of the system parameters values) denote the average, reference and demand power for the large scale water heater system. In particular, $P_{\text{dem}}[k] := P_{\text{ON}}[k] - P_{\text{OFF}}[k]$, where $P_{\text{ON}}[k]$ is the power drawn by all EWHs that are ON at time k and $P_{\text{OFF}}[k]$ is the power released by all EWHs that were ON at time $k-1$ and subsequently were switched OFF at time k. Given n_r in (1.14) and that PEM tracking is activated (per Fig. 1.4), the input $\beta[k]$ in Fig. 1.6b is designed, using information generated by the VPP's macro-model at each instant of time k, to be

$$\beta[k] = \min\left\{1, \frac{P_{\text{ref}}[k] - P_{\text{dem}}[k]}{P^{\text{rate}}\, n_r[k]}\right\}$$

when $P_{\text{ref}} > P_{\text{dem}}$ and 0, otherwise. Observe in the diagram that the timer dynamics automatically releases the population in $x_p^{n_p}$ and transitions them all to the OFF states. Also, note that if $\beta[k] = 0$ for all k then the state diagram becomes reducible since there the states cannot transition from ON to OFF. This last fact is undesirable given that x_{off}^1 ends up accumulating the entire population when k goes to infinity, which implies that every EWH becomes synchronized. This short-coming is addressed by additional states that will allow cold EWHs to turn ON even when the

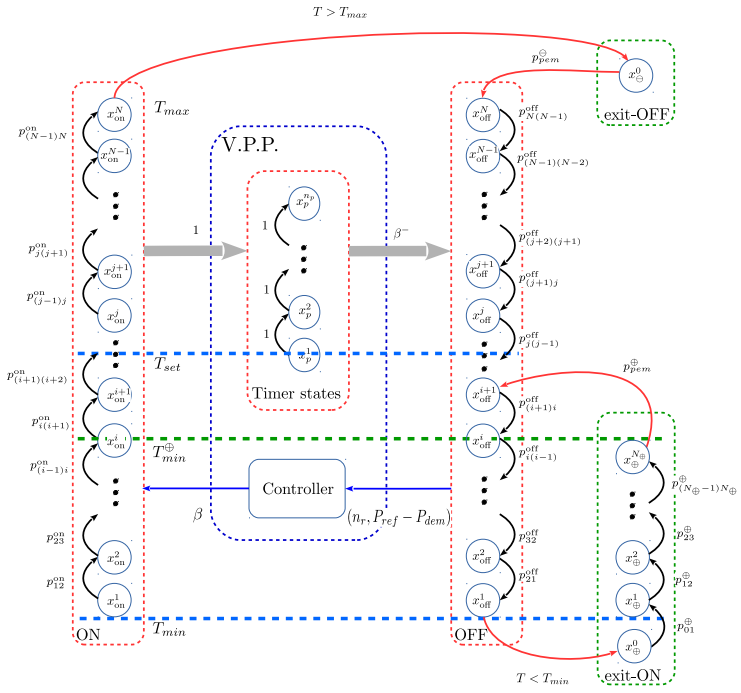

Fig. 1.7: PEM macro-model with exit-ON ($\oplus$) dynamics. ON/OFF state transitions are controlled by VPP and illustrated with gray and blue arrows.

VPP sets $\beta[k] = 0$. This *exit-ON/OFF* mechanism is augmented to the PEM macro-model to ensure QoS as described next.

Remark 1.4. The above design of input β is convenient, yet valuable, but can be improved by considering β in an optimal control setting, which is outside the scope of this chapter.

1.4.4 Exit-ON/OFF dynamics

As mentioned previously, the end-consumer QoS is of paramount importance when controlling a large scale system of DERs. Specifically, *no one likes to take a cold shower*. Therefore, whenever an EWH's temperature falls outside the dead-band $[T_{\min}, T_{\max}]$, it will exit the packetized scheme and revert to conventional control until a pre-specified PEM *exit-ON* set-point is reached. Once the *exit-ON* set-point is reached, the EWH is allowed to re-enter the packetized scheme.

The population of EWHs that are too cold and exit PEM (to turn ON) join the *exit-ON* mode dynamics (denoted by $\oplus$). On the other hand if a water heater is too hot and has to turn OFF, then it joins *exit-OFF* mode dynamics (denoted by $\ominus$) at state $x^0_\ominus$ after which EWHs transition under M naturally to the requesting states. These two PEM exit modes of operations were introduced in Section 1.3. Adding these modes of operation to the PEM macro-model only requires a simple augmentation of states with their corresponding transition rates as shown in Fig. 1.7. In the same figure, $T^\oplus_{\min} = T_{\text{set}} - T_{\text{PEM}}$ as explained in Section 1.3. The updated full population dynamics is given by (1.16) and

$$q[k+1] = M_{\text{exit}} \left(I + M^+_{\beta[k]} + M^-_{\beta^-[k]} \right) q[k], \quad y[k] = c^\top q[k],$$

where $M_{\text{exit}} := \text{diag}\{M_{\text{exit-ON}}, \bar{M}, M_{\text{exit-OFF}}\}$, $M_{\text{exit-ON}}$ is a matrix of zeros except for the main diagonal $(p^\oplus_{11}, \ldots, p^\oplus_{N_\oplus N_\oplus})$ and the first lower diagonal $(p^\oplus_{12}, p^\oplus_{23} \ldots, p^\oplus_{(N_\oplus - 1)N_\oplus})$, $M_{\text{exit-OFF}}$ introduces the probabilities to re-enter PEM from x^N_{on} to $x^0_\ominus$ and from x^N_{on} to x^N_{on} with $p^\ominus_{\text{pem}}$ corresponding to the *exit-OFF* mode, and $\bar{M}$ is such that $\bar{M}_{ij} = M_{ij}$ except for $M_{(N+N_\oplus+1)N_\oplus} = p^\oplus_{\text{pem}}$, which describe the transition probabilities to re-enter PEM from the *exit-ON* mode.

1.4.5 Validating the EWH macro-model against the micro-model

The internal temperature state and aggregate power output from a simulation of the macro-model is compared to those of a homogeneous agent-based (micro) simulation with $N = 1000$ packetized EWHs. Specifically,the parameters for the homogeneous collection of 1000 EWHs are chosen as: $P^{\text{rate}} = 4.5; T_{\text{inlet}} = T_{\text{amb}} = 14; L = 250; \eta = 1.0$. The simulation parameters are $\delta = 300$ secs, $\Delta t = 5$ secs The errors for the power signals are computed as

$$\text{E}_{\text{pow}} = \mathbb{E}[|y_{\text{micro}} - y_{\text{macro}^{\text{avg}}}| / y_{\text{micro}}] \times 100,$$

which is a percentage with respect to the micro-model simulation's output power. Also, the average temperatures profiles obtained from micro/macro-models are compared using $\text{RMSE}_{\text{temp}} = \sqrt{\mathbb{E}[(T^{\text{avg}}_{\text{micro}} - T^{\text{avg}}_{\text{macro}})^2]}$.

Fig. 1.8 is a 6 hour simulation of the homogeneous micro and macro models. Both simulations start by accepting all requests ($\beta = 1$) for the first two hours, which illustrates "control-free" PEM (i.e., only local stochastic access driving the system). After two hours, PEM tracking is enabled and it is observed that the average population temperature and aggregate power output of both simulations agree during most of the tracking period. Also, note that the tracking errors stay within $\pm 5\%$, which highlights PEM's ability to extract flexibility from the packetized population of loads and accurately track provided reference.

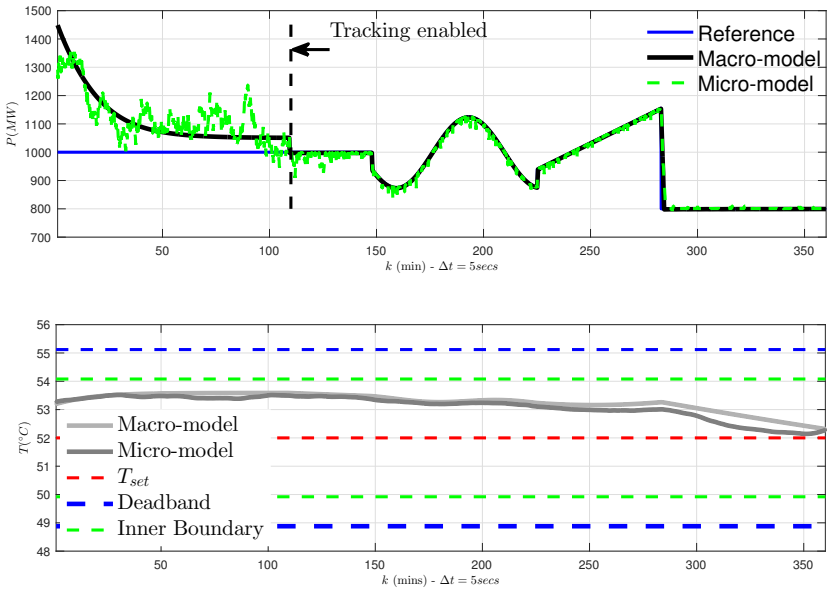

Fig. 1.8: Validating the homogeneous macro-model by comparing the aggregate power (top) and internal average temperature (bottom) against those of a homogeneous microsimulation with $N = 1000$ packetized electric water heaters.

Table 1.1: Capacity parameter for micro-model with $L = 250 + \sigma v$ and $v \sim N(0,1)$.

σ	E_{pow} ($\beta = 1$)	E_{pow}	$\text{RMSE}_{\text{temp}}$
10	1.5244	0.0646	0.2085
20	1.6054	0.0947	0.2118
30	1.7056	0.0699	0.2132
40	1.5968	0.0655	0.1960
50	1.6268	0.0619	0.2062

Table 1.2: Efficiency parameter for micro-model with $\eta = 1 - \sigma(1 - v)$ and $v \sim U(0,1)$.

σ	E_{pow} ($\beta = 1$)	E_{pow}	$\text{RMSE}_{\text{temp}}$
0.1	2.8933	0.1763	0.5685
0.2	6.7277	0.3715	0.9730
0.3	10.778	1.2613	1.3740
0.4	15.432	4.4017	1.7924
0.5	18.872	8.0183	2.1642

The accuracy of the macro-model under increasing levels of heterogeneity is shown in Tables 1.1 and 1.2 for a few salient parameters. Heterogeneity in parameters T_{amb} and T_{inlet} are omitted here as their impact on errors are similar to parameter L (Table 1.1). The efficiency parameter η has a strong effect on micro-simulation's aggregate power output and average temperature, however, it is expected that, in practice, η is uniform across the population (and close to 1.0). Overall, the macro-model is accurate and captures the dominant behavior of a large population of packetized EWHs, which makes it viable to be used for control and estimation to develop further insight into the capabilities of PEM.

1.5 Numerical case study of PEM with diverse DERs

While the previous section focused on homogeneous packetized TCLs, this section investigates how a single VPP, under PEM, can operate a diverse fleet of heterogeneous DERs. Specifically, the following case-study will illustrate how 1500 heterogeneous packetized TCL (1000), PEV (250), and ESS (250) devices can all be coordinated under with single VPP and simultaneously track a reference signal (in the aggregate) and satisfy (local) QoS constraints. For an in-depth case-study on just heterogeneous TCLs under PEM, please see [18], where a ramp-rate limit is introduced to the VPP to counteract the synchronization effects related to temperatures during long periods of reject-all.

The uncontrollable background demand for each load type describes the random perturbations to the local dynamic state.

- **TCL**: for the 1000 residential electric water heaters, the uncontrollable demand represents the use of hot-water in the home, such as a shower and running the washing machine or dishwasher. To model these hot-water events, we employ the following stochastic process:

  1. Choose the average number of hot-water (HW) events per hour, HW^{hr}_{avg}.
  2. For each TCL, uniformly distribute the total number of HW events with mean $2T_{sim}HW^{hr}_{avg}$.
  3. Randomly select hot-water events' starting times from available times, k^{HW}_0
  4. For each HW event, choose duration Δk^{HW} from normally distributed $\min\{\max\{N(700,300)/\Delta t,1\},3600/\Delta t\}$.
  5. From the duration of each HW event, choose a constant hot-water withdrawal rate $v_n[k]$ $[liters/min]$ based on the exponential distribution with mean $1200/(\Delta t\Delta k^{HW})$, which is inversely proportional to duration. A capacity of 30 liters/min is imposed on $v_n[k]$, which represents a high residential flow-rate [28].

- **PEV**: the background demand in the case of the 250 plug-in electric vehicle batteries represent the driving patterns that discharge the battery. The PEV travel patterns were randomly sampled from travel survey data [29] for New England, as described in [30], which provides the stochastic model for residential arrival and departure times, as well as miles driven. From an assumed electric driving range of 150 miles and an electric driving efficiency of 6.7 miles-per-kWh, the total reduction in SOC is computed upon arriving home (to charge).

- **ESS**: the 250 home batteries represent Tesla's PowerWalls (2.0), which have battery capacity of 13.5kWh, charge and discharge efficiency of around 95% (roundtrip of 92%), and a maximum (continuous) power rating of 5.0kW. It is assumed that the battery owner stochastically charges or discharges the battery based on a Gaussian random walk with a minimum power draw of 1.5kW in either direction. This could be representative of excess or deficit residential solar PV production or short-term islanding conditions.

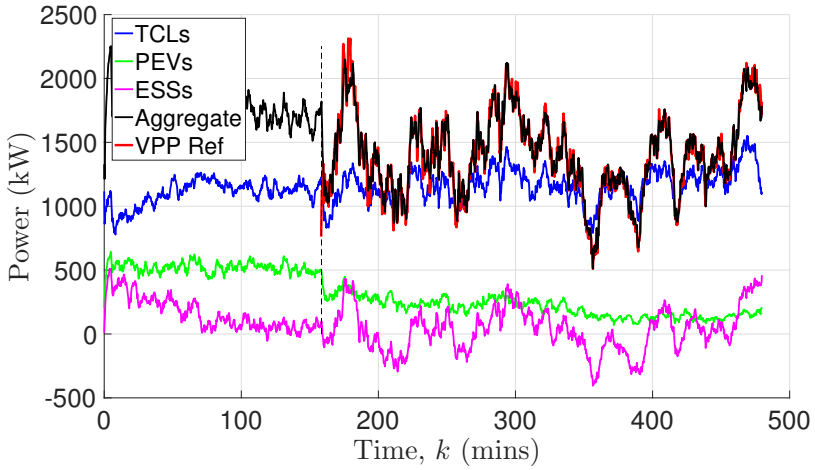

Fig. 1.9: The 8-hour case-study of a diverse VPP under PEM (with devices=1000 TCLs, 250 PEVs, and 250 ESS), where the initial time (0) represents 4:00PM while the end time is midnight, which affects the arrival/departure rates of the PEVs. The aggregate power produced by the VPP is shown for an initial accept-all phase and a later tracking phase (after minute 160). The aggregate VPP output power is shown and has mean average percent tracking error (MAPE) and RMSE of 2.03% and 4.51%, respectively. The aggregate power from each of the three DER types is provided as well. Note that the ESS devices operate about 0 kW, while the PEVs can charge only when home.

The $N = 1500$ diverse DER devices are then packetized and, over an 8-hour period (16:00 to 24:00), the VPP will interact with the loads and from 18:40 to 24:00 the VPP tracks a mean-reverting random signal that represents a balancing signal from the ISO. The tracking is achieved by denying or accepting packet requests based on real-time error between reference and aggregated VPP power output as described in Section 1.3. The tracking errors are less than 5% for packet epochs of $\delta = 5$ minutes. Figures 1.9 and 1.10 illustrate the tracking performance of the VPP and that QOS requirements are satisfied as well. Table 1.3 outlines the errors and other metrics of the diverse VPP during its tracking period, as a function of packet length. With increasing packet epoch δ, the flexibility of the VPP is reduced, which will increase tracking error metrics (MAPE, RMSE). However, with increasing packet lengths, the devices will cycle less often, which can help preserve the mechanical integrity of relays in electric water heaters (but may not have a significant effect on battery inverters). Therefore, there is a tradeoff between tracking performance and mechanical device degradation. In addition, as the packets become longer, the individual loads deviate further from their set-points, which implies that the VPP requires greater control effort despite the reduced tracking performance. Thus, there is a need to develop the analysis and optimal controller design for a VPP, which will be achieved by extending the Macro-model in Section 1.4 to include not just TCLs, but also PEVs and ESS. The final simulation further illustrates the value in managing a diverse fleet of DERs.

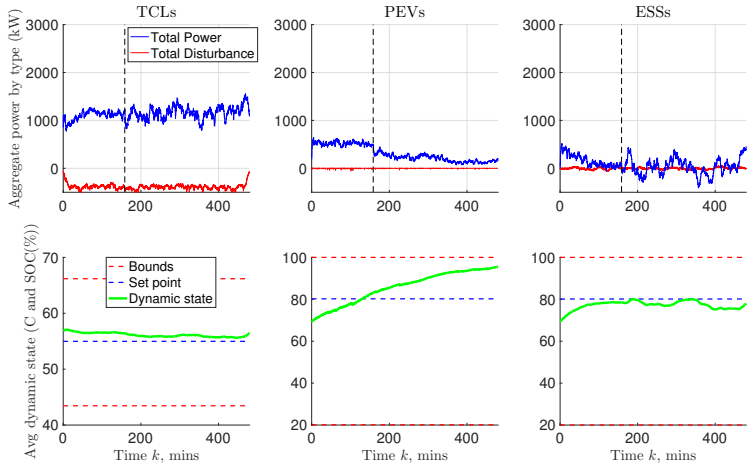

Fig. 1.10: Aggregate power and average dynamic state for each DER type. Despite the VPP tracking the reference for over five hours, the individual devices are able to both provide flexibility to VPP and satisfy QoS requirements. The left-most figures (top and bottom) are the TCLs, the center are PEVs, and the far-right are the ESS fleet. The TCL water heater and ESS battery should be close to their set-point, while the objective of a PEV is to a) cross set-point barrier and b) aim to be fully charged.

Table 1.3: Comparing tracking performance of diverse VPP for different packet lengths.

Metric	$\delta = 5$ mins	$\delta = 10$	$\delta = 15$	$\delta = 20$	$\delta = 30$
MAPE (%)	2.03	3.08	4.11	4.46	5.84
RMSE (%)	4.51	6.07	7.44	8.10	10.35
RMSE (kW)	58.93	76.76	86.93	98.13	126.01
Avg TCL ON/OFF cycles per hour	4.12	2.58	2.08	1.83	1.64
Avg device state deviation from set-point (%) [4]	4.69	4.96	5.28	5.51	5.91

Consider two VPPs: one is comprised of 500 TCLs, 250 PEVs, and 250 ESS batteries (i.e., diverse VPP) while the other contains 1000 TCLs (i.e., *TCL-only VPP*). Figure 1.11 illustrates how these two VPPs perform in tracking a signal composed of step, periodic, and ramp changes. It is clear that the diverse VPP out-performs the TCL-only VPP. In fact, the tracking RMSE for the diverse VPP is four times smaller than the TCL-only VPP (54kW vs. 220kW) while the MAPE is 30% lower (2% vs. 36%). Moreover, observe that this gain in performance comes without sacrificing QoS as the TCLs in both VPPs experience nearly identical mean absolute deviation from the temperature set-point: $2.4^{\circ}C$ vs. $2.5^{\circ}C$ (with similar standard deviations). To further illustrate the value of a diverse fleet of resources, Fig. 1.12 provides the ON/OFF statuses for each device in the respective VPPs. Careful comparison of the VPP illustrate that the TCL-only VPP fails to track the lower parts of the reference signal due to many TCLs opting out (i.e., transitions to *exit-ON*) as signified by very

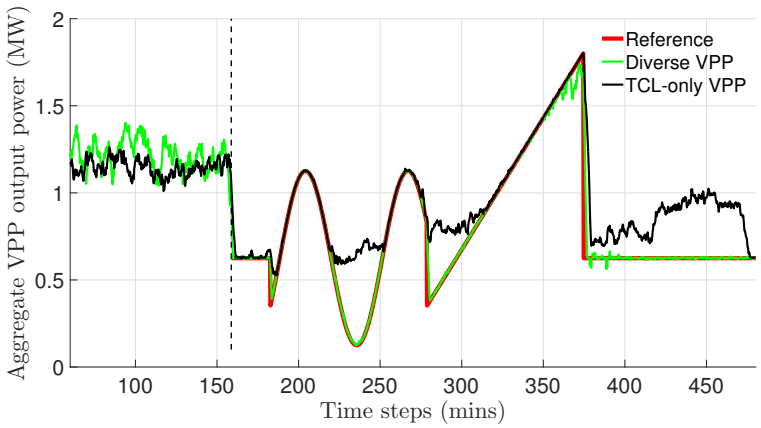

Fig. 1.11: Two VPPs are tracking a multi-mode reference signal with different sets of DERs. The diverse VPP (with 500 TCLs, 250 PEVs, and 250 ESS batteries) significantly outperforms the 1000 TCL-only VPP by leveraging the bidirectional capability of the batteries while maintaining QOS across all DER types. The TCL-only VPP is unable to track due to large number of TCLs the enter exit-ON and opt out of PEM.

long continuous ON periods for the TCL-only VPP in Fig. 1.12. That is, diversity in resources not only improves tracking ability, but also improves QoS delivered to end-consumer.

1.6 Conclusion and next steps for PEM

The focus of this chapter has been on the coordination of diverse DERs under PEM, including the validation of a macro-model for homogeneous TCLs and a case-study to support PEM under heterogeneity. These results lay the ground work for analysis to quantify macro-model uncertainty, forecast VPP flexibility capabilities and uncertainty bounds on performance, and enable the development of optimal control techniques for managing the VPP resources in order to improve tracking performance and QoS. Specifically, we are interested in embedding multiple VPPs into optimal power flow problems for transmission and/or distribution to develop grid-optimized reference signals that can be used in conjunction with real-time balancing between VPPs to improve resilience, reliability, and economics of power system operations.

This chapter has presented a novel and innovative paradigm for coordinating DERs: packetized energy management (PEM). PEM has numerous advantages over many of today's state-of-the-art coordination algorithms. At the core of PEM is local decision-making that randomizes the requests rates, which promotes asynchronous coordination across the population and protects the system from unwanted effects

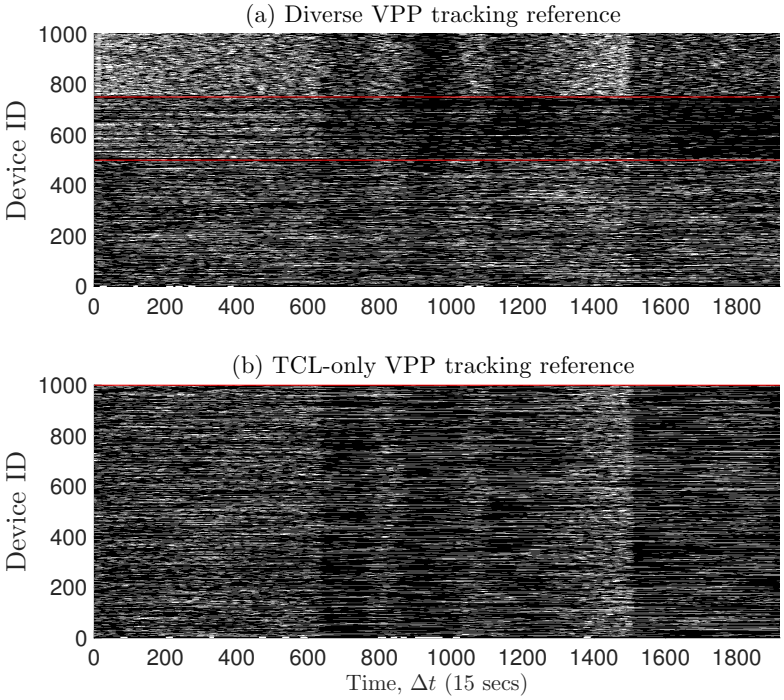

Fig. 1.12: The ON/OFF status of all devices in each VPP. The red lines indicate different DER types available to the diverse VPP in (a): 250 PEVs in the middle and 250 ESS batteries in the top band. The batteries are important during the downward phases of the signal as they offer the VPP added flexibility. In addition, the diversity helps avoid large-scale opting out of TCLs after the final step-change (in minute 380).

of synchronization. That is, PEM is truly a bottom-up approach that protects the privacy of the consumer. In addition, since a VPP only requires a measurement of the aggregate power output and the packet request rate, PEM offers a simple framework for modeling and control that avoids having to rely on entire histograms to be transmitted to devices for control.

These unique properties of PEM have been explored in this chapter through simulations and modeling. On the modeling front, a state-bin transition population model is augmented with a timer that tracks the completion of a packet, and informs the VPP of expiring packets (i.e., device that soon will switch OFF). This information will be valuable as we build up the optimal control framework for PEM. In addition, the model captures the opt-out mechanism, which provides a mechanism for improving QOS (albeit it by reducing available flexibility).

Acknowledgements This work was supported by the U.S. Department of Energys Advanced Research Projects Agency - Energy (ARPA-E) grant DE-AR0000694.

References

[1] J. Frolik, "Qos control for random access wireless sensor networks," in *Proc. 2004 Wireless Communications and Networking Conference (WCNC04)*, Greenville, SC, 2004.

[2] N Abramson, "Throughput of packet broadcasting channels," *IEEE Transactions on Communications*, vol. 25, no. 1, pp. 117–128, Jan. 1977.

[3] G Bianchi, "IEEE 802.11-saturation throughput analysis," *IEEE Communications Letters*, vol. 2, no. 12, pp. 318 –320, Dec. 1998.

[4] D. S. Callaway and I. A. Hiskens, "Achieving Controllability of Electric Loads," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 184–199, Jan. 2011.

[5] M. Morgan and S. Talukdar, "Electric power load management: Some technical, economic, regulatory and social issues," *Proceedings of the IEEE*, vol. 67, no. 2, pp. 241 –312, 1979.

[6] F. C. Schweppe, R. D. Tabors, J. L. Kirtley, H. R. Outhred, F. H. Pickel, and A. J. Cox, "Homeostatic utility control," *IEEE Trans. Power Apparatus and Systems*, no. 3, pp. 1151–1163, 1980.

[7] J. L. Mathieu, S. Koch, and D. S. Callaway, "State Estimation and Control of Electric Loads to Manage Real-Time Energy Imbalance," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 430–440, 2013.

[8] Y. Chen, A. Busic, and S. P. Meyn, "State estimation and mean field control with application to demand dispatch," in *2015 54th IEEE Conference on Decision and Control*, IEEE, 2015, pp. 6548–6555.

[9] W. Zhang, J. Lian, C.-Y. Chang, and K. Kalsi, "Aggregated Modeling and Control of Air Conditioning Loads for Demand Response," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4655–4664, 2013.

[10] S. Esmaeil Zadeh Soudjani and A. Abate, "Aggregation and Control of Populations of Thermostatically Controlled Loads by Formal Abstractions," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 3, pp. 975–990, May 2015.

[11] 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. 99, pp. 1–11, Jul. 2016.

[12] S. P. Meyn, P. Barooah, A. Busic, Y. Chen, and J. Ehren, "Ancillary Service to the Grid Using Intelligent Deferrable Loads," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2847–2862, Nov. 2015.

[13] B. Zhang and J. Baillieul, "A packetized direct load control mechanism for demand side management," in *IEEE Conference on Decision and Control*, 2012, pp. 3658–3665.

[14] B. Zhang and J. baillieul, "A novel packet switching framework with binary information in demand side management," in *IEEE Conference on Decision and Control*, IEEE, 2013, pp. 4957–4963.

[15] B. Zhang and J. Baillieul, "Control and Communication Protocols Based on Packetized Direct Load Control in Smart Building Microgrids," *Proceedings of the IEEE*, vol. 104, no. 4, pp. 837–857, 2016.

[16] J. Frolik and P. Hines, "Urgency-driven, plug-in electric vehicle charging," in *Proc. IEEE PES Innovative Smart Grid Technol. (ISGT-Europe)*, Berlin, 2012.

[17] P. Rezaei, J. Frolik, and P. D. H. Hines, "Packetized Plug-in Electric Vehicle Charge Management," *IEEE Transactions on Smart Grid*, pp. 1–10, Nov. 2013.

[18] M. Almassalkhi, J. Frolik, and P. Hines, "Packetized energy management: asynchronous and anonymous coordination of thermostatically controlled loads," *American Control Conference*, 2017.

[19] L. Duffaut Espinosa, M. Almassalkhi, P. Hines, S. Heydari, and J. Frolik, "Towards a Macromodel for Packetized Energy Management of Resistive Water Heaters," *IEEE Conference on Information Sciences and Systems*, 2017.

[20] J. Frolik and P. Hines, "Random access, electric vehicle charge management," in *1st IEEE International Electric Vehicle Conference. (IEVC)*, Greenville, 2012.

[21] P. Rezaei, J. Frolik, and P. Hines, "Packetized plug-in electric vehicle charge management," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 642–650, 2014.

[22] C. Goh and J Apt, "Consumer strategies for controlling electric water heaters under dynamic pricing," *Proceeding of Carnegie Mellon Electricity Industry Center*, 2004.

[23] J. Kondoh, N. Lu, and D. J. Hammerstrom, "An Evaluation of the Water Heater Load Potential for Providing Regulation Service," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1309–1316, 2011.

[24] N Lu and D. P. Chassin, "A state-queueing model of thermostatically controlled appliances," *IEEE Transactions on Power Systems*, vol. 19, no. 3, pp. 1666–1673, Aug. 2004.

[25] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice Hall, 1986.

[26] S. G. Buchberger and L Wu, "Model for instantaneous residential water demands," *Journal of Hydraulic Engineering*, 1995.

[27] H. Zhang, W. S. Gray, and O. R. Gonzalez, "Performance Analysis of Digital Flight Control Systems With Rollback Error Recovery Subject to Simulated Neutron-Induced Upsets," *IEEE Transactions on Control Systems Technology*, vol. 16, no. 1, pp. 46–59, 2008.

[28] ASHRAE, "Chapter 49: Service Water Heating," in *ASHRAE Applications Handbook*, ASHRAE, Oct. 2002, pp. 49.1–49.22.

[29] U.S Department of Transportation.

[30] A. D. Hilshey, P Rezaei, P Hines, and J Frolik, "Electric Vehicle Charging: Transformer Impacts and Smart, Decentralized solutions," in *IEEE Power and Energy Society General Meeting*, IEEE, 2012, pp. 1–8.