Reference-Tracking Control Policies for Packetized Coordination of Heterogeneous DER Populations

L. A. Duffaut Espinosa⬤, *Senior Member, IEEE*, Adil Khurram, *Student Member, IEEE*, and Mads Almassalkhi⬤, *Senior Member, IEEE*

Abstract—This manuscript presents design and analysis of a set of reference-tracking control policies for large-scale coordination of distributed energy resources (DERs) and quantifies tracking errors that arise due to heterogeneity in the power ratings for a fleet of DERs. In particular, the relay-based, reference-tracking control strategy that underpins much of packetized energy management (PEM) is augmented to uniquely leverage PEM's energy packet request mechanism to optimize the number of accepted requests and to explicitly consider the quality of service (QoS). In addition, tracking errors from modeling a heterogeneous fleet of packetized DERs with a group of homogeneous macromodels are analytically derived for relevant PEM information scenarios. Finally, simulation-based analysis validates the results and shows that PEM is suitable for providing load balancing and ramping services for the grid.

Index Terms—Demand dispatch, distributed control, distributed energy resources (DERs), packetized energy management (PEM), virtual power plant (VPP).

I. INTRODUCTION

BALANCING demand and supply is a fundamental problem in power system operations. For decades, the operating paradigm has been *supply follows demand*, which has been implemented via the hierarchical primary, secondary, and tertiary frequency regulation schemes. However, with increasing penetrations of variable renewable generation, such as solar PV and wind, and internet-enabled, connected, and controllable appliances (e.g., distributed energy resources or DERs), it is now technically and economically feasible for flexible demand to follow a variable supply (i.e., provide dynamic grid services) [1]–[3]. However, to control large fleets of DERs to provide grid services requires a distributed controller implementation, which has been known since the late 1970s [4], [5]. More recently, the notions of demand dispatch [6] and controllable loads [7] have reinvigorated the field of controlling DERs. In fact, it has become evident that large-scale coordination of DERs will require randomization to avoid harmful effects of synchronization [8]–[11]. Thus, this article focuses on so-called distributed, randomized control policies for demand dispatch that embed control logic both at the DER (device) and at the coordinator, which is also known as *aggregator* or *virtual power plant* (VPP). Information is then exchanged between the DERs and the coordinator, which informs the device to probabilistically transition its operational state (e.g., consume, standby, and/or supply) and permits the coordinator to compute control inputs regulating the controllable net demand (total consumption minus variable supply).

However, coordinating large fleets of DERs is challenging and requires careful distributed controller design that is cognizant of one or more of the following: 1) the devices' finite energy limits (e.g., temperature bounds for thermostatically controlled loads or TCLs); 2) ON/OFF cycling rates; 3) communication costs; 4) fleet heterogeneity within and between device classes; and 5) privacy concerns around data sharing. It is within this context that a set of reference-tracking controllers are presented here for a new demand dispatch method called packetized energy management (PEM). PEM can coordinate diverse DER populations within a single ("coupled") controller to track an aggregate power regulation signal, while guaranteeing local DER-specific QoS constraints, and is described as follows.

1) A DER estimates its local need for energy, e.g., energy state of charge or SOC.
2) If the SOC is within a predefined range of comfort, the DER probabilistically requests from the coordinator to consume energy at a fixed rate (e.g., 4 kW) for a pre-specified epoch (e.g., 5 min), which begets an *energy packet* (e.g., 0.33 kWh). If the SOC falls below a pre-defined lower threshold the DER automatically and temporarily opts out of PEM to guarantee QoS and reverts to a conventional control mode (e.g., charges) until the SOC is returned within limits and after which it returns to PEM operation. Similarly, the DER also opts out if the SOC exceeds a pre-specified higher threshold.
3) The asynchronous coordinator or VPP either accepts or denies the DER's packet request on an arrival basis in relation to the current reference tracking error. If the request is denied, go to *i*); else, consume the energy packet and then go to *i*).

The packetizing concepts that inspired PEM focused on coordinated EV charging under a fixed transformer constraint [12], [13]. Recently, PEM has been adapted for coordination of water heaters [14] together with batteries [15] and a

macro-level (aggregate) model has been characterized qualitatively and quantitatively in [16]–[18]. Most recently, Duffaut Espinosa and Almassalkhi [19] provides the most complete description of the macromodel and QoS requirements. In this article, the focus is entirely on design of control policies for the PEM aggregator that dynamically optimizes the number of accepted requests and the QoS under different information scenarios. However, PEM is not the first article to explore DER control and coordination at scale.

Methods based on mean-field models and control provided an early approach for coordinating large-scale population of DERs, including TCLs [20], [21]. The seminal work in [10] designed a feedback linearization scheme around a state-bin transition model where the aggregate power tracking error input together with estimated bin states were used to compute an updated state-bin transition probability vector that was broadcast to all devices. The feedback linearization overcomes a state-input bi-linearity in the controller, which requires that an accurate estimate of the states is available to compute the control input. The work herein also includes bi-linear input-state term, however, the control signal for PEM is a scalar signal that represents the proportion of available devices to transition (i.e., communicated only to a small sub-set of available devices). Mathieu's work has been advanced to consider higher-order TCL models [22], multiple market objectives with a virtual energy storage model [23], and communication delays [24].

The approach in [11] also considers a randomized control policy for a homogeneous population of DERs, but avoids a state-bin transition model by instead using a mean-field model based on a linearization around the average device's pre-determined baseline transition matrix. The linearization enables the use of a single scalar that can be broadcast to update the shape of the local transition probabilities. Recently, the group has been active in extending the work to consider estimation of the QoS (e.g., a proxy for state of charge and on–off device cycling rates) with and without local opt-out control [25], [26], intra-class DER heterogeneity via filtering [27], and virtual energy storage models [28]. Related work on stochastic control and modeling of heterogeneous TCLs can be found in [29] where the control strategy combines stochastic transitions with temperature dead-band (hysteresis) control to provide primary and secondary frequency regulation services for the grid. While the aggregate (macro) model in our work is a state-bin transition model similar to that in Mathieu, PEM's approach replaces the controllable probability-to-transition curve at the flexible load with a fixed probability-to-request curve that engenders a stochastic request process input to the PEM coordinator. The control strategy then is to determine the (scalar) proportion of requests to accept, which can be implemented as a probabilistic response to requesting devices or via recent Internet-of-Things (IoT) methods as a binary (accept/deny) response to each device. PEM has used the latter binary approach in earlier work on PEVs and TCLs [13], [14].

PEM's device-driven request-response mechanism was inspired by early TCP/IP and ALOHA network protocols. Similar approaches can be found in [30], which presents a decentralized packetized coordination scheme that is analyzed via queueing theory to determine QoS under a fixed, known resource budget. The hierarchical controller proposed in [31] is similar in spirit to PEM, but considers a non-randomized, priority-based scheme wherein a device regularly transmits a transition fitness score (e.g., in [0, 1]) instead of a request, which the coordinator uses to rank and prioritize the individual devices for control. In addition (unlike PEM), the controller uses the ranking to form on-to-off and off-to-on transition priority queues from which they can construct aggregate up/down flexibility curves that are used to transmit and update device-level set-points to ensure a rapid autonomous frequency response. This device-driven scheme is well suited for smaller populations due to the device-by-device prioritization and regular information exchanges. Another distributed DER control scheme replaces the randomized device-level transition with a deterministic, cyclical decision process for set-point control of TCLs [32]. Other related works include [33] where the effects of dynamic constraints and modeling errors are quantified.

Thus, the article herein first proposes control architectures for diverse fleets of PEM enabled DERs by leveraging the author's prior work on a PEM macromodel [16]–[19]. These architectures allow the coordinator to either generate a common control signal for different DER-types with the objective to track a market reference signal, or to obtain a distinct control signal for each type of DER that simultaneously tracks the aforementioned reference as well as satisfies a secondary objective such as QoS guarantees, minimize number of accepted requests etc. Then, the synchronous nature of the macromodel is analytically characterized and compared with the asynchronous request-response mechanism in PEM. Finally, a group-based approach to handle DER intra-class heterogeneity is provided as well as bounds on tracking error. The specific contributions of this article are as follows:

1) Two different DER coordination architectures are proposed and validated for coordinating diverse fleets of DERs. In the first architecture, denoted the *coupled architecture*, each packet request identifies as either a charge request or a discharge request only, that is the VPP is unaware of the DER-type the request is coming from. In this case, the VPP accepts the same proportion of requests across all DER classes. The second architecture is the *decoupled architecture* in which the DER-type is included in the packet request. The VPP then optimizes charge and discharge acceptance rates for each DER class separately.

2) Novel control policies are designed for tracking reference power signals that are applicable to both coupled and decoupled architectures. These policies have as secondary objectives the maximization or minimization of the number of packet requests to be accepted (input optimization) and guaranteeing QoS close to a specific energy state (state optimization). One can further co-optimize requests (input) and QoS (state), which bears resemblance to optimal controller problems. With respect to several classes of DERs, the control policies are validated for both architectures and performance metrics are provided.
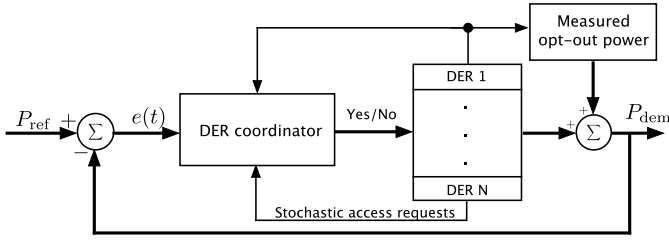
Fig. 1. Closed-loop feedback system for PEM with P_{ref} provided by the grid or market operator and the aggregate net-load P_{dem} measured by the VPP.

3) Analytical characterization of the relationship between the controllers implemented in PEM's asynchronous request-response mechanism and the corresponding synchronous representation in the aggregated bin transition macromodel that describes the mean-field behavior. This type of characterization is useful when implementing the control schemes herein under coupled and decoupled architectures in cloud-based serverless implementations of PEM.

4) Analysis of parametric heterogeneity unique to PEM's packet request-response mechanism is conducted by studying uncertainty in the estimated rated power of requests. The tracking performance of a large group of DERs subject to parametric heterogeneity is compared against g smaller groups of DERs with the same rated power of request and modeled by the macromodel described in Section II-B. Analytical bounds on tracking errors are then calculated for the g groups.

This article is organized as follows. Section II summarizes PEM and the macro-model. In Section III, reference-tracking controllers that maximize/minimize the acceptance rate of packet requests is presented along with a controller that maintains QoS. In Section IV, parametric heterogeneity with respect to DERs' rated power is discussed and the worst-case tracking error under steady-state conditions is presented. Conclusions and future research directions are given in the final section.

II. MODEL FOR PACKETIZED ENERGY MANAGEMENT

This section provides context for the main results of this article and summarizes the work in[14], [17], and [19].

A. Packetized Energy Management

For each DER under PEM, quality of service (QoS) is ensured by utilizing a probabilistic request rule based on the local dynamic state. In addition, if the QoS falls below a certain user-defined threshold, then the DER is capable of exiting PEM temporarily. Synchronization is mitigated by introducing randomization in the DERs' request rates based on the local state variables. The typical closed loop block diagram of a PEM system is shown in Fig. 1.

Let the energy state of the nth DER in a fleet be denoted as z_n. The equation for this DER is given by the following discrete-time dynamic model

$$z_n[k+1] = f_n\big(z_n[k], \phi_n[k], P_{c,n}^{\text{rate}}, P_{d,n}^{\text{rate}}, w_n[k]\big) \qquad (1)$$

where f_n is a 1-D mapping (usually linear or bilinear). Energy transfer rates of the nth DER are given by $P_{c,n}^{\text{rate}}$ and $P_{d,n}^{\text{rate}}$ when charging (c) and discharging (d), respectively. The hybrid state ϕ_n corresponds to the set of modes $\{c, \text{sb}, d\}$ associated with {charge, standby, discharge}, respectively [14], [17]. In this article, the focus is on coordinating electric water heaters (EWHs), bidirectional energy storage systems (ESSs), and electric vehicles (EVs). The EWHs and EVs are seen by aggregators as having charging dynamics only whereas ESSs corresponds to both charging and discharging dynamics. The parameter $w_n \in \mathbb{R}$ maps the end-use consumption to the energy state and is explained next for each DER class considered in this article.

The EWH end-use consumption w_n corresponds to the hot water extraction from the tank. Since water extraction always leads to loss of heat energy, therefore, $w_n \leq 0$. In this article, Poisson random pulses are used to model the end-user process as detailed in [17], [19], and [35].Unless otherwise stated, the EWH parameters used in this article are the same as in [19, Tab. 1]. The ESS units are assumed to be installed for the purpose of backup power by the end-user and are managed by the aggregator (e.g., Tesla or VPP herein) to provide grid services. This setup is based on the state of Vermont's largest utility, Green Mountain Power, and their residential Tesla Powerwall battery program [35]. In this program a customer can pay \$15/month per ESS (5 kW/13 kWh). This program has been popular as many rural customers opt for multi-day backup capability with two ESSs and, today, Vermont has more than 2000 PowerWalls under control. Therefore, it is assumed in all the simulations presented in this work, the ESS are rated at 5 kW/13.5 kWh, charge and discharge efficiency of around 95% (roundtrip of 92%) along with the set-point of 70% [19] and the background usage is absent. Finally, the driving patterns of the EV driver are modeled using a two-state Markov chain consisting of two modes; *driving* and *parking* [19]. Note that in [19], *standby* was used instead of *parking*. This driving pattern then determines the end-user consumption of EVs considered in this work. The average drive time of the EV driver is chosen to be approximately 30 min and the EVs are assumed to have an electric driving range of 150 miles with the driving efficiency of 7 miles-per-kWh [19]. The request probability in PEM is presented next.

The probability that the nth DER with dynamic state $z_n[k] \in [\underline{z}_n, \overline{z}_n]$ and desired set-point $z_n^{\text{set}} \in (\underline{z}_n, \overline{z}_n)$ over time k (for discretization time-step Δt) makes a request is given by a cumulative distribution function over the range of admissible dynamics states. For this purpose, an exponential distribution encoding three desirable conditions has been chosen; $i)$ no request (i.e., $\Pr(z_n) = 0$) from device n for $z_n \geq \overline{z}_n$; $ii)$ guaranteed request (i.e., $\Pr(z_n) = 1$) from device n for $z_n \leq \underline{z}_n$; and $iii)$ $\Pr(z_n) = p_R$ when $z_n \equiv z_n^{\text{set}}$ for a designed mean-time-to-request value, $1/p_R$, at the device's desired set-point, z_n^{set}, during interval Δt. This design is aimed at attracting a device's state towards the set-point. The probability of request is then given by

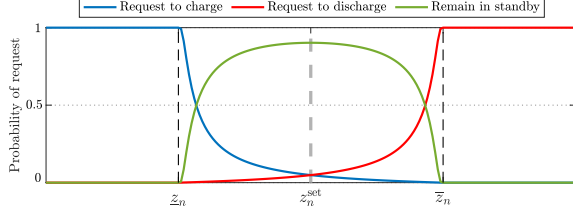$$\Pr(z_n[k]) := 1 - e^{-\mu(z_n[k])\Delta t}$$

Fig. 2. Probability of request curves showing the effect of local state z_n on the packet request probabilities. The blue line corresponds to (2), the red line corresponds to discharging packets, and the green line gives the complementary probability of remaining in standby state. Here, $m_R = 1/300$ Hz and $\Delta t = 15$ seconds.

where $\mu(z_n[k]) > 0$ is a variable rate parameter dependent on the local dynamic state. For *charging* energy packet requests

$$\mu(z_n[k]) = \begin{cases} 0, & \text{if } z_n[k] \geq \overline{z}_n \\ m_R \left(\dfrac{\overline{z}_n - z_n[k]}{z_n[k] - \underline{z}_n} \right) \cdot \left(\dfrac{z_n^{\text{set}} - \underline{z}_n}{\overline{z}_n - z_n^{\text{set}}} \right), & \text{if } z_n[k] \in (\underline{z}_n, \overline{z}_n) \quad (2) \\ \infty, & \text{if } z_n[k] \leq \underline{z}_n \end{cases}$$

where $m_R > 0$ [Hz] is a design parameter that defines the mean time-to-request (MTTR) for $z_n = z_n^{\text{set}}$ which is fixed at the mid-point of $(\underline{z}_n, \overline{z}_n)$. A similar expression follows for $\mu(z_n[k])$ in the case of *discharging* packets. Fig. 2 shows the request probability curves for charging and discharging packets together with the probability of remaining in standby during coordination. While (2) is one possible choice for probability of request, however, this design can be changed or incorporated in the control (as done in [10] and [11]) and is the topic of ongoing work.

B. PEM Macro-Model

The state bin transition macro-model for a large homogeneous population of DERs *of the same load type* is presented next. A macro-model for a diverse population is comprised of a finite number of different type DER homogeneous populations that work together under the same VPP for demand dispatch. Hence, consider a population of DERs described by (1) with common underlying state space. To create a finite-state abstraction of the entire population evolution, the state space is discretized in a manner that the main features of the system are preserved. Furthermore, the size of the population is chosen so that the law of large numbers holds and the system behavior is approximated by a single averaged effect. This means that the macromodel is robust against the effects of individual DERs as the number of devices increases [36], [37]. The PEM macromodel is introduced next.

The transition probabilities between bins are determined from the dynamical system equations of the DERs comprising the population with respect to the type of end-user events for the particular class of DERs [19]. Let $\bar{\mathcal{X}} = \{x_1, \ldots, x_N\}$, where each element is called a state and constitute a consecutively ordered partitioning of the continuous state space Z in which the DERs evolve. Denote by q_j the probability of being in state x_j and $q := (q_1, \ldots, q_N)^T$ the probability distribution over $\bar{\mathcal{X}}$. For the particular case of DERs having hybrid one

dimensional dynamics as in (1), an interval $[z_{\min}, z_{\max}]$ within Z is divided into N consecutive bins each corresponding to a *bin state* in $\bar{\mathcal{X}}$, where $x_i \in \bar{\mathcal{X}}$ corresponds to the interval $[z_{i-1}, z_i) \subset [z_{\min}, z_{\max}]$. Since (1) includes three types of dynamics (charge/standby/discharge), the state space for the system consists of the union of the three identical copies of $\bar{\mathcal{X}}$ given by $\mathcal{X} = \mathcal{X}_c \cup \mathcal{X}_{\text{sb}} \cup \mathcal{X}_d$. At time k, the probability mass function of the system is $q^\top = (q_c^\top, q_{\text{sb}}^\top, q_d^\top)$ with $q_c = (q_c^1, \ldots, q_c^N)^\top$ and q_{sb} and q_d defined similarly. Note that q contains the percentage of the population in each state of $\mathcal{X}$ so that, $\mathbf{1}_{3N}^\top q = 1$, where, $\mathbf{1}_{3N}^\top \in \mathbb{R}^{3N}$ is a vector of all ones. Therefore, if N_e is the total number of DERs and $N_{e,c}^i$ is the number of devices in state x_c^i, then $N_{e,c}^i = q_c^i N_e$. Similarly, the percentage of N_e that is charging and discharging, and the total power demand of the system due to charging and discharging packets are

$$y_c = C_c q, \quad y_d = C_d q, \quad \text{and} \quad y_{\text{dem}} = C_{\text{dem}} q \quad (3)$$

where $C_c = (\mathbf{1}_N^\top, 0 \cdots 0) \in \mathbb{R}^{3N}$, $C_d = (0 \cdots 0, \mathbf{1}_N^\top) \in \mathbb{R}^{3N}$, $C_{\text{dem}} = N_e (P_c^{\text{rate}} C_c - P_d^{\text{rate}} C_d) \in \mathbb{R}^{3N}$, $\mathbf{1}_N = (1, \ldots, 1)^\top \in \mathbb{R}^N$. In this article, symmetric charging and discharging is assumed, that is, P_c^{rate} and P_d^{rate} are equal. In case these are not the same, their average is taken instead for the sake of simplicity. Henceforth, the subscript is dropped and P^{rate} is used for both charge and discharge rated powers. Furthermore, let $T_{\text{req,c}}^i$ ($T_{\text{req,d}}^i$) be the probabilities associated with making a charge (discharge) request using (2) in state x_{sb}^i. Then the probability of requesting a charging packet and not a discharging packet is given by $T_{\text{req,c},d̸}^i = T_{\text{req,c}}^i (1 - T_{\text{req,d}}^i)$ [17]. Similarly, one can obtain $T_{\text{req,d},c̸}^i = T_{\text{req,d}}^i (1 - T_{\text{req,c}}^i)$. Furthermore, it should be noted that simultaneous charging and discharging requests can occur with probability $T_{\text{req,c,d}}^i = T_{\text{req,c}}^i T_{\text{req,d}}^i$. However, no request is submitted in that case. That is, making a charge or discharge request are mutually exclusive events. The total number of charging and discharging requests are

$$y_{\text{req,c},d̸} = C_{\text{req,c},d̸} q, \quad \text{and} \quad y_{\text{req,d},c̸} = C_{\text{req,d},c̸} q$$

where $C_{\text{req, c,} d̸} = (\mathbf{0}_N^\top, \mathcal{T}_{\text{req, c,} d̸}^\top, \mathbf{0}_N^\top) \in \mathbb{R}^{3N}$, $C_{\text{req, d,} c̸} = (\mathbf{0}_N^\top, \mathcal{T}_{\text{req, d,} c̸}^\top, \mathbf{0}_N^\top) \in \mathbb{R}^{3N}$, $\mathcal{T}_{\text{req, c,} d̸} := (T_{\text{req, c,} d̸}^1, \ldots, T_{\text{req, c,} d̸}^N)$, $\mathcal{T}_{\text{req, d,} c̸} := (T_{\text{req, d,} c̸}^1, \ldots, T_{\text{req, d,} c̸}^N)$. In Section IV, uncertainty in P^{rate} due to heterogeneity in the DER population's rated power is introduced and analyzed.

The discrete-time equation modeling the population dynamics of a PEM system is given by

$$q[k+1] = M \tilde{M}_{\beta, \beta_{\text{sb}}}[k] q[k] \quad (4)$$

and $y[k] = (y_{\text{dem}}[k], y_{\text{req, c,} d̸}[k], y_{\text{req, d,} c̸}[k]) \in \mathbb{R}^3$, where the aggregate net-load (P_{dem}) is obtained from $y[k]$ as $P_{\text{dem}}[k] = (1, 0, 0) y[k]$. Furthermore, the 2-D control signal $\beta := (\beta_c, \beta_d)$ defines the proportion of charge/discharge packet requests out of the total charge/discharge requests denoted as n_r^c / n_r^d, that are accepted (i.e., defines how many DERs transition from standby (sb) to charge (c) and discharge (d)). It should be noted here that $\beta_c \in [0, 1]$ and $\beta_d \in [0, 1]$ since the total number of accepted charging and discharging requests cannot exceed n_r^c and n_r^d, respectively. In addition,

$\beta_{sb} := (\beta_c^-, \beta_d^-)$ is the proportion of DERs in (c) and (d), respectively, whose energy packets end now and return to (sb). The matrices in (4) (omitting time dependence for all βs) have the form

$$M = \left(\begin{array}{c|c|c} M_c & M_{c,sb} & 0_N \\ \hline M_{sb,c} & M_{sb} & M_{d,sb} \\ \hline 0_N & M_{sb,d} & M_d \end{array} \right) \text{ and}$$

$$\tilde{M}_{\beta,\beta_{sb}}$$

$$:= \left(\begin{array}{c|c|c} (1-\beta_c^-)I_N & \beta_c T_{req,c,\&} & 0_N \\ \hline \beta_c^- I_N & I_N - \beta_c T_{req,c,\&} - \beta_d T_{req,d,\&} & \beta_d^- I_N \\ \hline 0_N & \beta_d T_{req,d,\&} & (1-\beta_d^-)I_N \end{array} \right)$$

where 0_N denotes the N-dimensional zero matrix, I_N denotes the N-dimensional identity matrix, M_h, for $h = \{c, sb, d\}$ is a multi-diagonal matrix containing the probabilities of staying, going to higher energy states and going to lower energy states, $M_{c,sb}$ and $M_{sb,c}$ are responsible for transferring any DER that exceeds z_{max} from c to sb and any DER that falls behind z_{min} from sb to c, respectively, and similarly $M_{d,sb}$ and $M_{sb,d}$ provide the transition probabilities from d to sb and from sb to d including the probabilities of uncontrollable events, $T_{req,c,\&} = \text{diag}\{\mathcal{T}_{req,c,\&}\}$ and $T_{req,d,\&} = \text{diag}\{\mathcal{T}_{req,d,\&}\}$. The computation of the transition probabilities has been addressed in [17] and [19].

Modeling the evolution of the number of active DER charging and discharging packets in the system requires introducing two sets of timer states (c and d). That is, given packet epoch δ, the sampling time step Δt, and two timer states vectors $x_{p,h} \in \mathbb{R}^{n_p}$ with $n_p = \lfloor \delta/\Delta t \rfloor$ and $h = \{c, d\}$, the *timer dynamics* are given by

$$x_{p,h}[k+1] = M_{p,h}x_{p,h}[k] + B_{p,h}q_h^+[k] \qquad (5)$$

where $q_h^+ := \beta_h \bar{q}_{sb}^h$, $\bar{q}_{sb}^h := T_{req,h}q_{sb}$, $T_{req,c} = T_{req,c,\&}$, $T_{req,d} = T_{req,\&,d}$ and $B_{p,h} \in \mathbb{R}^{n_p \times N}$ is responsible for allocating the new charge/discharge population into their corresponding charge/discharge timer states. The number of charging/discharging packet requests received by the VPP is then $n_r^h := \mathbf{1}_N^\top \bar{q}_{sb}^h$ with $h = \{c, d\}$ and $\mathbf{1}_N = (1, \ldots, 1)^\top \in \mathbb{R}^N$. The timer provides the formula for the percentage of DERs whose packet expires. That is, $\beta_h^- := x_{p,h}^{(n_p)} / \sum_{i=1}^{n_p} x_{p,h}^{(i)}$, where $x_{p,h}^{(i)}$ is the ith component of $x_{p,h}$.

C. Opt-Out Dynamics

Finally, the opt-out control mechanism that ensures QoS for each DER under PEM is introduced. This is achieved by augmenting q in (4) with *opt-out* states $q_\oplus$. That is, q is redefined as $q^\top := (q_\oplus^\top, q^\top)$ and

$$q[k+1] = M_{exit}\tilde{M}_{\beta,\beta_{sb}}[k]q[k] \qquad (6)$$

where $\tilde{M}$ is trivially augmented with a diagonal block identity matrix and zeros everywhere else (leaving the states q_{opt} unaffected by β and β_{sb})

$$M_{exit} := \left(\begin{array}{c|c} M^\oplus & M_{pem}^\ominus \\ \hline M_{pem}^\oplus & M \end{array} \right)$$

$M^\oplus$ is a submatrix (minor) of M that has all rows and columns corresponding to states higher than the prespecified PEM
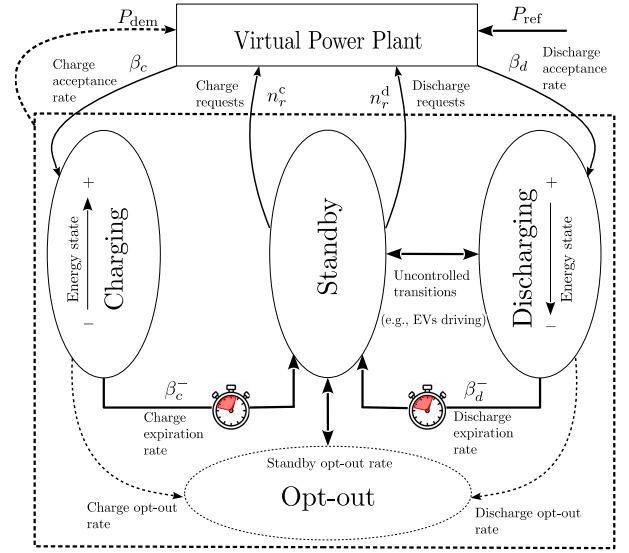

Fig. 3. Transition diagram of a DER population under PEM. In the case of EVs, the transition from standby to discharging is uncontrolled. The opt-out state ensures QoS and the timers capture packet completion rates.

re-entry bound removed, $M_{pem}^\ominus$ provides the transition probabilities of leaving PEM and $M_{pem}^\oplus$ provide the transition probabilities of reentry PEM. Fig. 3 depicts a complete diagram of a DER population under PEM. In this diagram, controlled transitions among charging, standby, and discharging states are regulated by the VPP whereas uncontrolled transitions[1] are also permitted. For a detailed discussion of QoS guarantees, see [19].

Including opt-out dynamics completes the PEM macromodel that captures the aggregate dynamics of a large enough population of DERs. Theoretically, there exist modeling errors as a result of the finite-state abstraction [33]. However, the intended use of the macromodel in this work is to design reference tracking control policies for specific objectives such as satisfying QoS. In this regard, empirical evidence is provided here that shows that the chosen population size is sufficient for the Law of Large Numbers to hold. For that purpose and without loss of generality, consider a population of EWHs when all requests are accepted, that is, $\beta_c = 1$. In this situation, all DERs get what they need without imposing restrictions from the coordinator. The error in the total power demand between the PEM macromodel and simulated EWH fleet in steady state, called *modeling error in power* in this work, is compared for different population sizes. The mean of the modeling error in power was observed to be close to zero whereas the standard deviation decreases with the size of the population as expected. Fig. 4 shows the standard deviation normalized with respect to the size of the population and the rated power. It can be seen that the standard deviation decreases proportionally to $1/\sqrt{N}$, where N is the size of the population. The normalized standard deviation of the modeling error for 250 DERs is approximately 2.5% that corresponds

---

[1]The transition of some flexible loads, such as EVs, from Standby to Discharge is not controllable and occurs at a rate based on average driving behaviors without the intervention of the VPP.
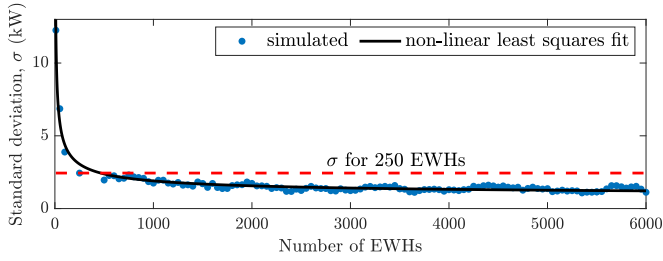
Fig. 4. Standard deviation of the model error normalized with respect to the number of EWHs and the rated power is plotted against the population size. Also, shown here is the least-square fit to $f(N) = \alpha_1/\sqrt{N} + \alpha_2$ for the population size N and $(\alpha_1, \alpha_2) = (36.55, 0.75)$. Matlab was used to fit $f(N)$ to the standard deviation and the goodness of the fit is given by $R^2 = 0.97$ indicating a good match.
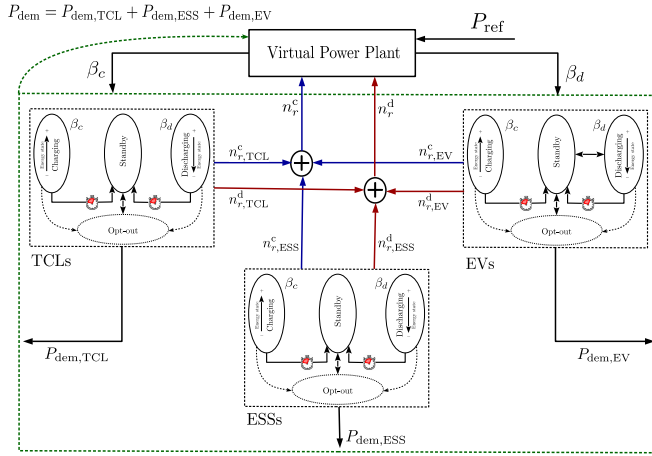

Fig. 5. Diverse system of TCLs, ESSs, and EVs for a single *coupled* PEM coordinator. The VPP sees only the aggregated power demand as well as the total incoming charging and discharging request rates. A decoupled PEM implementation involves the PEM coordinator receiving total requests from each DER class independently and generating a control signal for each of such classes of DERs.

to ± 6.25 EWHs and for 5000 DERs is approximately 1.2% amounting to ± 60 EWHs, which is an acceptable tolerance given the fact that the coordinator has no information about the DERs state. Therefore, in the simulations presented in this article, the population consists of a minimum of 250 DERs of the same type. Moreover, the effect of modeling errors on the design of control policies presented in this work will be explored in future publications.

D. Control Architectures for Diverse DERs

While coordinating different classes of DERs within a fleet, the type of DER (e.g. TCL, ESS etc.) information can be included in the packet request along with the size of the packet (e.g. 4.5 kW). This gives rise to two VPP architectures, namely, coupled and decoupled. If the DER type is included in the request, then the VPP generates a separate control input for each class within the fleet. Whereas, in the absence of DER type information, the VPP generates a single control input for the entire fleet. The coupled implementation is illustrated in Fig. 5 for a fleet consisting of TCLs, ESSs, and EVs.

III. REFERENCE-TRACKING CONTROL POLICIES FOR PEM

This section presents novel reference-tracking control policies for PEM that explicitly considers the aggregate packet request rates and QoS. Specifically, control policies are implemented that optimize the number of accepted requests and preserves the QoS. This represents an extension from the request-maximizing policy presented in [17] and [18]. However, in order to track a reference while satisfying QoS for a population, a *power baseline* or *nominal response* for PEM needs to be established first.

A. Nominal Response of DER Fleet Under PEM

The flexibility available to a VPP for tracking a grid or market signal is subject to QoS constraints on the population of DERs under PEM. In particular, there exists a constant reference signal (or nominal demand), which maintains the average SOC above the desired value (to preserve QoS), and above/below which the average SOC increases/decreases. That is, the nominal demand represents a simple sustainable trajectory and produces the VPP's nominal response. A definition of the nominal response of a PEM system is provided and discussed next.

Definition 1: The nominal response of a fleet of DERs under PEM, given by (6), is the minimum constant power signal for which QoS is sufficiently satisfied for the average DER.

This nominal response provides a reference about which flexibility can be measured and establishes a relationship between aggregate VPP demand, QoS, and average SOC, which is the first step in the development of a dispatchable virtual energy storage model [23], [38]. This nominal response is characterized in PEM by the *nominal control* $\beta^* = (\beta_c^*, \beta_d^*)$ that is the solution of the following constrained optimization problem

$$\beta_c^*, \beta_d^* = \underset{\beta_c, \beta_d \in [0,1]}{\arg\min} \sum_{i=1}^{n} C_{\mathrm{dem},i} q_i^* \quad \text{s.t.} \tag{7a}$$

$$q_i^* = M\tilde{M}(\beta, \beta_{\mathrm{sb}}) q_i^*, \tag{7b}$$

$$(q_i^*)^\top x_v^i \geq z_{\mathrm{set}}^i \tag{7c}$$

where $C_{\mathrm{dem},i} q_i^* = P_i$ is the total packetized nominal demand from the ith class of DERs at steady state q_i^*. The vector $x_v^i \in \mathcal{X}^i$ contains the parameters associated with the discretized bin values of the ith DER class (e.g., temperatures, SOCs, etc.) and z_{set}^i is the desired set point for the dynamic state of the ith DER class.

1) For TCLs and EVs, the optimal solution to (7) is attained for $\beta_c \in [0, 1]$ due to $\sum_{i=1}^{n} C_{\mathrm{dem},i} q_i^*$ being a monotonically increasing function of β_c (i.e., less packets accepted is less power consumed) and $(q_i^*)^\top x_v^i$ in (7c) is also monotone in β_c (i.e., accepting less requests leads to a lower average SOC, z_{set}^i). However, including the ESS class introduces $\beta_d \in [0, 1]$ to regulate the acceptance rate of discharging packet requests from batteries that seek to discharge. This additional degree of freedom means that multiple ESS solutions can achieve the same net-power (charge minus discharge)
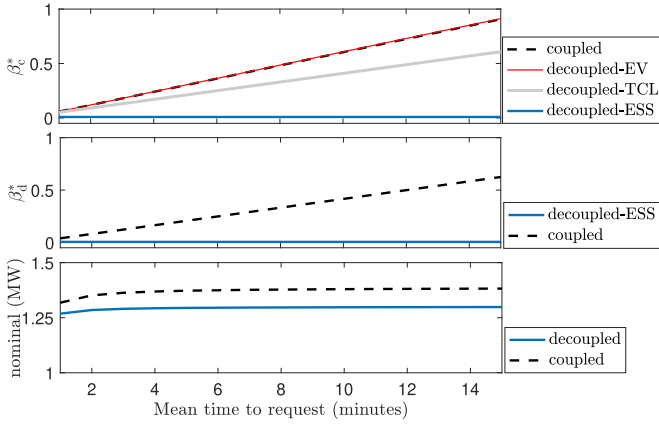
Fig. 6. Effect of changing m_R on nominal control and power is illustrated here for both decoupled and coupled implementations. (Top) Acceptance rate of charging packets β_c^*, (center) acceptance rate of discharging packets β_d^*, (bottom) nominal power consumption of the fleet.
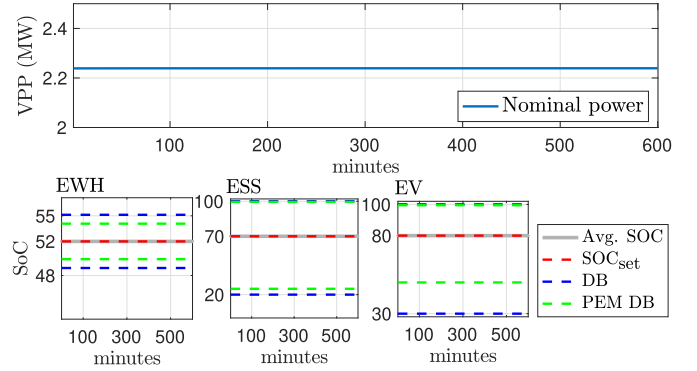

Fig. 7. Nominal response for a decoupled implementation of PEM maintains the average QoS for a VPP with 1000 TCLs, 1000 ESSs and 250 EVs. The mean dynamic state of each DER population satisfies (7c) with equality for the specified population set points.

without affecting the ESS average SOC. Hence, with the ESS class, the optimal solution is not unique. Work is ongoing to formally prove and characterize the optimal solution under various implementations of PEM and the scope herein is meant to introduce properties of the nominal response to motivate the reference tracking controller developed in the next section.

2) When a PEM system is in equilibrium (power and SOC) due to a fixed control input β, the internal signal β_{sb} becomes time invariant, and packet interruptions are then negligible since the average SOC is near the set-point as shown in [18]. Specifically, it was also shown that $\beta_{sb} = (\beta_c^-, \beta_d^-) \approx (1/n_p, 1/n_p)$ when tracking admissible trajectories.

3) In the case of coupled implementation (7) generates β^* for the entire fleet (TCLs, ESSs and EVs), that is, a single nominal β^* is obtained for all DER types.

4) The optimization problem presented in (7) is clearly non-convex, however, ongoing efforts consider convex re-formulations. However, those results are outside the scope of this manuscript.

5) The nominal control input β^* also depends upon the parameter m_R in the probability of request. Recall that m_R is a design parameter that defines the meantime to request. Consider first the nominal charging acceptance rate β_c^*. Fig. 6 shows that increasing m_R results in an increase in β_c^* in both decoupled and coupled implementations. This makes sense because increasing m_R leads to DERs requesting less often. Therefore, the VPP has to accept more requests to maintain the fleet at a nominal steady state. Similarly, β_d^* also increases with m_R. It should be noted that the change in β_c^* and β_d^* for ESS is insignificant since the background usage is absent. However, the nominal power consumption of the fleet remains approximately constant. This study provides empirical evidence that the amount of power required by the diverse DER fleet to operate at nominal conditions remains the same for $m_R > 2$ min. The VPP, on the other hand, requires larger control effort when m_R is increased. Similar behavior is observed for nominal

discharge acceptance rate β_d^*. Nevertheless, further work is required for fully characterizing the behavior of m_R and its effect on control effort that will be the focus of future publications. All simulations in this manuscript uses m_R equal to the packet length.

Fig. 7 shows the nominal response of a VPP with three DER populations (TCLs, ESSs, and EVs). The top plot in Fig. 7 shows the aggregate power demand of the fleet and the bottom plots shows the state of charge. Furthermore, in Fig. 7 and in subsequent figures, the blue dotted lines represent the deadband (DB) and the green dotted lines resrepsent the PEM deadband (PEM DB). This simulation shows the nominal control input for TCLs to be $\beta = (0.210, 0)$, ESSs have $\beta = (0.009, 0.006)$ and EVs have $\beta = (0.302, 0)$. Note that the components of β for ESSs are close to zero since the background noise process is ignored (i.e., batteries are used only to provide flexibility). For EVs introduced in [19], an average driving time of 30 min and a driving occupancy of 20% was used. The nominal control for the coupled implementation yields $\beta = (0.302, 0.208)$ that amounts to average TCL temperature, ESSs SOC, and EV SOC of 52.4 °C, 70% and 80%, respectively. Note that EVs in this manuscript can only request charging packets and do not participate in vehicle-to-grid (V2G) discharging events. This means that the average SOC for a population EVs can only decrease during periods when the population drives more than it charges (on average). The main takeaway from Fig. 7 is that when the fleet starts in a nominal steady state, then the diverse DER fleet maintains its SOC under nominal acceptance rate, β^*.

B. Reference-Tracking Controller With Diverse DERs

While the previous section focused on nominal control of a constant trajectory, this section provides explicit control policies for real-time reference-tracking by VPP. In fact, herein a diverse VPP is illustrated, i.e., a single coordinator for homogeneous groups of DERs. The control law that achieves the desired reference tracking determines the proportion of charging and discharging requests that are accepted (i.e., $\beta = (\beta_c, \beta_d)$) and is straightforward: 1) measure the tracking error and 2) accept requests so as to minimize the error. This

logic is similar in effect to ACCEPT/DENY relay control for the individual packet requests. However, at the macro-level with an incoming stochastic stream of requests, the control effectively becomes continuous in [0, 1] where zero (0) means that all requests are denied and one (1) implies all requests are accepted.

Note that since each packet request can also include its corresponding charge/discharge energy transfer rate, e.g., $P_{h,n}^{\mathrm{rate}}$, which provides the macromodel with an accurate estimate of average packet height $P_h^{\mathrm{rate}} := 1/\beta_h n_r^h \sum_{n=1}^{\beta_h n_h} P_{h,n}^{\mathrm{rate}}$ for $h = \{c, d\}$. Let P_{ref} be the reference power signal provided by some grid or market condition and let P_{dem} be the actual VPP net-load (charging minus discharging packets). In a coupled implementation of PEM, the two scalar control inputs β_c and β_d are generated by the VPP so that these minimize the tracking error. However, unlike other DER coordination schemes, PEM's unique packet request mechanism allows the VPP to consider maximizing or minimizing the number of accepted requests without affecting the aggregate VPP net-demand.

The scheme is shown in Fig. 5 for a diverse DER fleet consisting of TCLs, ESSs, and EVs. Despite the fact that there is only one VPP for all the DER populations, the VPP can generate either one control signal $\beta = (\beta_c, \beta_d)$ for all the DER populations (*coupled* case) or multiple control signals designed for the specific DER population (*decoupled* case). The latter will obviously require more information (e.g., a request could include pinging the VPP the DER type). In the coupled case, for the purpose of tracking, the VPP solves the following optimization problem at time-step k:

$$\chi_c^*, \chi_d^* = \underset{\substack{\chi_c \in [0, n_r^c[k]] \\ \chi_d \in [0, n_r^d[k]]}}{\arg\min} \ (\mathcal{F}(\chi)[k])^2 \qquad (8)$$

where $\mathcal{F}(\chi)[k] := \chi_c P_c^{\mathrm{rate}}[k] - \chi_d P_d^{\mathrm{rate}}[k] - \epsilon[k]$ and $\epsilon[k] := P_{\mathrm{ref}}[k] - P_{\mathrm{dem}}[k] - n_{\mathrm{sb},c}[k]P_c^{\mathrm{rate}} + n_{\mathrm{sb},d}[k]P_d^{\mathrm{rate}}$ is the tracking error of the VPP that includes the effect of expiring packets in terms of $n_{\mathrm{sb},c}$ and $n_{\mathrm{sb},d}$ which is the number of DERs that move from charge and discharge modes to standby modes respectively. Thus, after solving (8) the VPP controls are given by $\beta_c[k] = \chi_c^*/n_r^c[k]$ and $\beta_d[k] = \chi_d^*/n_r^d[k]$. Observe that $(\mathcal{F}(\chi))^2$ is only positive semi-definite so the optimal solution is not unique as formulated above. Thus, one can add a secondary objective to maximize/minimize the total number of accepted packet requests (i.e., $\chi_c^* + \chi_d^*$). The solution to (8) is provided by the theorem below.

Theorem 1: Define the available packet budget as $n_{\mathrm{error}} = \epsilon/P_c^{\mathrm{rate}}$ when $\epsilon \geq 0$ and $n_{\mathrm{error}} = |\epsilon|/P_d^{\mathrm{rate}}$ otherwise. The solution (χ_c^*, χ_d^*) to (8) that maximizes and minimizes the number of accepted requests is given in Tables I and II, respectively.

Proof: Due to the simplicity of the proof, the focus will be in the case where there are enough requests to track the signal and for maximizing number of requests and $\epsilon > 0$. The other cases follow directly. When there are enough requests to track the reference signal, it follows that $\mathcal{F}(\chi) = 0$. Then, $\chi_d = (P_c^{\mathrm{rate}}\chi_c - \epsilon)/P_d^{\mathrm{rate}} \in [0, n_r^d]$. This solution corresponds to a family of solutions for $\chi_c \geq \epsilon/P_c^{\mathrm{rate}}$ since discharging

TABLE I
MAXIMIZING THE NUMBER OF ACCEPTED REQUESTS

Cases		$\alpha_r > n_{\mathrm{error}}$	$\alpha_r \leq n_{\mathrm{error}}$
$\epsilon > 0$ $(\alpha_r := n_r^c)$	$\chi_c^*:$	$\min\{n_r^c, (n_d^d P_d^{\mathrm{rate}} + \epsilon)/P_c^{\mathrm{rate}}\}$	n_r^c
	$\chi_d^*:$	$\min\{n_r^d, (n_r^c P_c^{\mathrm{rate}} - \epsilon)/P_d^{\mathrm{rate}}\}$	0
$\epsilon \leq 0$ $(\alpha_r := n_r^d)$	$\chi_c^*:$	$\min\{n_r^c, (n_r^d P_d^{\mathrm{rate}} - \epsilon)/P_c^{\mathrm{rate}}\}$	0
	$\chi_d^*:$	$\min\{n_r^d, (n_r^c P_c^{\mathrm{rate}} + \epsilon)/P_d^{\mathrm{rate}}\}$	n_r^d

TABLE II
MINIMIZING THE NUMBER OF ACCEPTED REQUESTS

Cases		$\alpha_r > n_{\mathrm{error}}$	$\alpha_r \leq n_{\mathrm{error}}$
$\epsilon > 0$ $(\alpha_r := n_r^c)$	$\chi_c^*:$	n_{error}	n_r^c
	$\chi_d^*:$	0	0
$\epsilon \leq 0$ $(\alpha_r := n_r^d)$	$\chi_c^*:$	0	0
	$\chi_d^*:$	n_{error}	n_r^d

requests can cancel charging requests and thus increase χ_c. Therefore, if $(P_c^{\mathrm{rate}}\chi_c - \epsilon)/P_d^{\mathrm{rate}} \leq n_{ri}^d$, one can set $\chi_c = n_r^c$ and $\chi_d = (P_c^{\mathrm{rate}}n_r^c - \epsilon)/P_d^{\mathrm{rate}}$ when this number is less than n_r^d, otherwise $\chi_d = n_r^d$. On the other hand, if $(P_c^{\mathrm{rate}}\chi_c - \epsilon)/P_d^{\mathrm{rate}} > n_r^d$, then $\chi_d = n_r^d$ and $\chi_c = (P_d^{\mathrm{rate}}n_r^d + \epsilon)/P_c^{\mathrm{rate}} > 0$ when this number is less than n_r^c, otherwise $\chi_c = n_r^c$. In both cases, the total number of accepted requests is maximized. This completes the case shown in Table I for $\epsilon > 0$ and $\alpha_r := n_r^c > n_{\mathrm{error}}$. The case for $n_r^c \leq n_{\mathrm{error}}$ is obtained by obeying the constraints on the number of accepted requests in the problem. ∎

The mechanics for the decoupled case are similar to the coupled one, except that n_r^h for $h = \{c, d\}$ is no longer the total number of requests across all DER classes, but the number of requests coming from each DER class separately. This decoupled case implies an implicit disaggregation of the reference signal at each time instant. The work in [39] also deals with the disaggregation of the reference signal in the frequency domain so that DER types are pre-selected to track the corresponding pieces. This reduces tracking to considering individual DER types tracking their own reference independently of each other. The decoupled case here relies on (8) and a secondary objective (e.g., maximizing or minimizing requests in Theorem 1, which dissaggregates the reference at each time step. In this case, if suddenly some DER types become unavailable (e.g., TCL peak time of the day or EVs start driving en masse due to the end of the working day) then the other DER populations automatically begin to track the part of the reference that was left unattended by the offline DERs. In other words, the decoupled case performs a dynamic disaggregation of the reference. However, when dealing with DER classes with different rated powers, decoupled implementation can potentially prioritize one class of DERs over the others. Consider for example two DER classes with rated powers 4 and 5 kW, tracking a common reference signal while maximizing the total number of accepted requests. The

decoupled VPP is inclined towards accepting more requests from 4 kW DER class than the class with rated power 5 kW since the objective is to accept as many requests as possible. It is worth pointing out here that although the grid access of the 5-kW class is reduced, VPP's main objective of tracking the reference signal is not compromised. Therefore, to avoid prioritization of 4-kW DER-class over the other, the objective function in (9) can be regularized by penalizing accepting requests from the 4 kW DER class. Another approach involves a QoS-optimizing policy that prioritizes access to the grid based on the energy needs of the DER class.

A QoS-optimizing control policy for accepting packet requests seeks to track a reference signal while limiting the deviation of the average SOC from the desired set-points. Compared with the two previously described Minimize/Maximize requests policies, the QoS-optimizing control policy consists of QoS (state) and number of accepted requests (input) that resembles an optimal controller problem [28]. The decoupled QoS-optimizing control policy aims to find the optimal solution, $\chi_{c,d}^{\mathrm{QoS}}$, such that

$$\chi_{c,d}^{\mathrm{QoS}}[k] = \underset{\substack{\mathcal{F}(\chi)[k]=0 \\ 0 \leq \chi_{c,d} \leq n_r^{c,d}[k]}}{\arg\min} \quad \mathrm{QoS}_{\chi_{c,d}}[k] \qquad (9)$$

where the objective function is given by

$$\begin{aligned} &\mathrm{QoS}_{\chi_{c,d}}[k] \\ &= \nu_{\mathrm{TCL}}^{\mathrm{QoS}} \left| z_{\chi_{c,d}}^{\mathrm{TCL}}[k] - z_{\mathrm{set}}^{\mathrm{TCL}} \right|^2 + \nu_{\mathrm{ESS}}^{\mathrm{QoS}} \left| z_{\chi_{c,d}}^{\mathrm{ESS}}[k] - z_{\mathrm{set}}^{\mathrm{ESS}} \right|^2 \\ &\quad + \nu_{\mathrm{EV}}^{\mathrm{QoS}} \left| z_{\chi_{c,d}}^{\mathrm{EV}}[k] - z_{\mathrm{set}}^{\mathrm{EV}} \right|^2 + \nu_{\mathrm{nom}}^{\top} \left| \chi_{c,d}[k] - \chi_{c,d}^{\mathrm{nom}}[k] \right|^2 \quad (10) \end{aligned}$$

$z_{\chi_{c,d}}^{\mathrm{TCL}}[k]$, $z_{\chi_{c,d}}^{\mathrm{ESS}}[k]$ and $z_{\chi_{c,d}}^{\mathrm{EV}}[k]$ denote the average dynamic state of TCLs, ESSs, and EVs at time k as a function of $\chi_{c,d} = (\chi_{\mathrm{TCL,c}}, \chi_{\mathrm{ESS,c}}, \chi_{\mathrm{ESS,d}}, \chi_{\mathrm{EV,c}})^{\top}$ and $n_r^{c,d} = (n_{r,\mathrm{TCL}}^c, n_{r,\mathrm{ESS}}^c, n_{r,\mathrm{ESS}}^d, n_{r,\mathrm{EV}}^c)^{\top}$ is the vector of charging and discharging requests corresponding to TCLs, ESSs, and EVs respectively. Furthermore, $\chi_{c,d}^{\mathrm{nom}} = \beta_{c,d}^{\mathrm{nom}\top} n_r^{c,d}$ where $\beta_{c,d}^{\mathrm{nom}} := (\beta_{\mathrm{TCL,c}}^{\mathrm{nom}}, \beta_{\mathrm{ESS,c}}^{\mathrm{nom}}, \beta_{\mathrm{ESS,d}}^{\mathrm{nom}}, \beta_{\mathrm{EV,c}}^{\mathrm{nom}})^{\top}$. The latter is obtained from the nominal problem of (7) for the decoupled VPP case. Finally, $\nu_{\mathrm{nom}} := (\nu_{\mathrm{TCL,c}}^{\mathrm{nom}}, \nu_{\mathrm{ESS,c}}^{\mathrm{nom}}, \nu_{\mathrm{ESS,d}}^{\mathrm{nom}}, \nu_{\mathrm{EV,c}}^{\mathrm{nom}})^{\top}$ and the scalars ν represent weights on the different objectives. For the coupled VPP case, the box constraint in (9) reduces to $0 \leq \chi_c \leq n_r^c$ and $0 \leq \chi_d \leq n_r^d$, where n_r^c and n_r^d are, respectively, the total charging and discharging requests of all the DER classes combined. Naturally, one requires to know the current state distribution q of the respective populations. In [18], an extended Kalman filter (EKF) was constructed for the state-bin transition model providing the dynamics of DERs populations. The same EKF formulation is employed to estimate $\mathrm{QoS}_{\mathrm{TCL}}[k]$, $\mathrm{QoS}_{\mathrm{ESS}}[k]$ and $\mathrm{QoS}_{\mathrm{EV}}[k]$ at time k in a manner that the distance between the average dynamic state at k and a pre-established set point is minimized by $\chi_{c,d}^{\mathrm{QoS}}$. It is important to highlight here that the scheme is not manipulating individual DER's SOC but choosing a control input β such that the aggregated SOC behaves as desired.

Figs. 8–10 illustrate the decoupled control policies for maximizing, minimizing, and QoS-optimizing the number of accepted packet requests for a diverse VPP with 1000
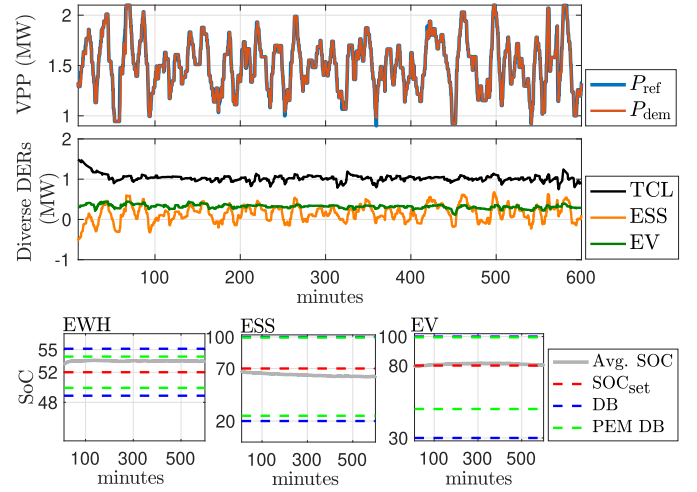

Fig. 8. Decoupled maximizing policy on a fleet of 1000 TCLs, 1000 ESSs, and 250 EVs. Top plot shows power at the VPP for reference tracking. Middle plot gives the individual power from each DER-type macromodel. Bottom plots give the SOC for each individual macromodel.
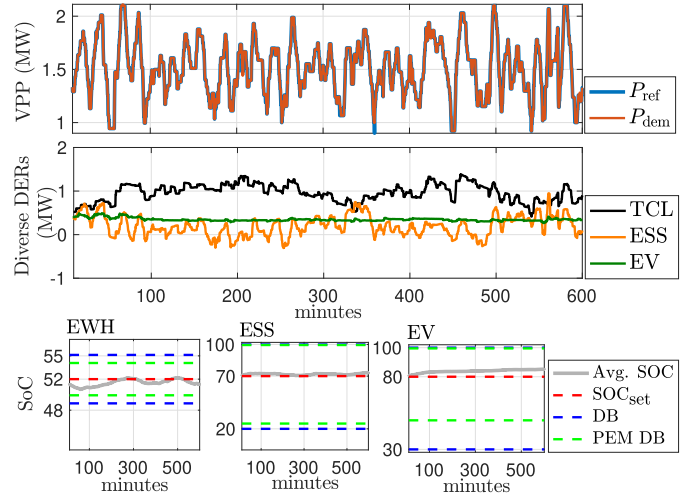

Fig. 9. Decoupled minimizing policy on a fleet of 1000 TCLs, 1000 ESSs, and 250 EVs. Top plot shows power at the VPP for reference tracking. Middle plot gives the individual power from each DER type macromodel. Bottom plots give the SOC for each individual macromodel. Note that the EVs' decrease in SOC is due to more cars driving (and discharging) than charging.

TCLs, 1000 ESSs, and 250 EVs. The reference signal in the simulations represents a de-trended (with respect to the nominal response) and scaled AGC signal obtained from [40] and shows that the VPP is able to provide approximately ± 0.5 MW of flexibility. Furthermore, it should be noted that in Fig. 9 the average power consumption of ESSs is greater than zero where the average SOC remains close to the setpoint. This is because of the nonunity battery efficiency [19] These optimization problems were solved using `fmincon` in Matlab on a MacBook Pro with a 2.2-GHz processor and 16-GB memory and the solution was obtained on average within 45 ms.

1) Reference Tracking Performance: The root mean square (RMS) tracking error between the reference signal and the total power output is reported in Table III for the coupled and decoupled implementations of PEM. The one-step-ahead
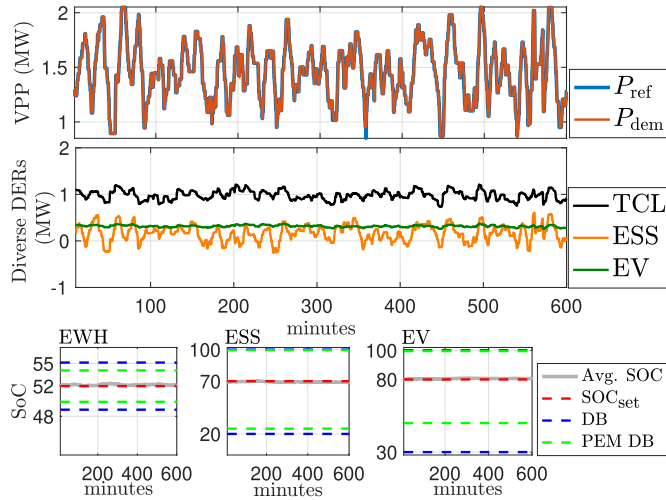
Fig. 10. Decoupled QoS policy on a fleet of 1000 TCLs, 1000 ESSs, and 250 EVs. Top plot: power at the VPP for reference tracking. Middle plot: individual power from each DER type macromodel. Bottom plots: SOC for each individual macromodel.

up (down) flexibility for a single DER population with n_r^c charging requests and n_r^d discharging requests specifies the maximum increase (decrease) in power demand P_{dem} due to that DER population. Specifically, it is defined as follows:

$$\overline{flex}[k] = (n_r^c[k] - n_{\text{sb},c}[k])P_c^{\text{rate}} + n_{\text{sb},d}[k]P_d^{\text{rate}}$$
$$+ (n_\oplus[k] - n_\oplus^{\text{sb}}[k])P^{\text{rate},\oplus},$$
$$\underline{flex}[k] = (n_r^d[k] - n_{\text{sb},d}[k])P_d^{\text{rate}} + n_{\text{sb},c}[k]P_c^{\text{rate}}$$
$$- (n_\oplus[k] - n_\oplus^{\text{sb}}[k])P^{\text{rate},\oplus}$$

where $n_{\text{sb},h}$, for $h = c, d$ is defined in Section III-B, $n_\oplus$ is the number of DERs opting out, $n_\oplus^{\text{sb}}$ the number of DERs leaving opt-out mode to sb mode and $P^{\text{rate},\oplus}$ is the average power rate of DERs in opt-out mode. In the decoupled implementation, the total upward (downward) flexibility is obtained as the sum of all individual DER classes upward (downward) flexibilities within a fleet. The coupled implementation, however, considers the aggregated charge and discharge requests, aggregated charging and discharging expiring packets as well as the aggregated number of DERs leaving and entering the opt-out mode to compute the up (down) flexibility using the average P_c^{rate} and P_d^{rate} rates within the DER classes in the fleet. The system flexibility is a function of β in that $\overline{flex}[k]$ corresponds to $\beta_c = 1$ and $\beta_d = 0$ and $\underline{flex}[k]$ corresponds to $\beta_c = 0$ and $\beta_d = 1$. In addition to $\overline{\text{RMS}}$ tracking error, the average of the one-step-ahead up/down flexibility is provided in Table III. The RMS tracking error and one-step-ahead flexibility suggest that minimizing the number of accepted requests improves flexibility – the QoS policy is similar too and was implemented with $(\nu_{\text{TCL}}\ \nu_{\text{ESS}},\ \nu_{\text{EV}}) = (19, 1, 1)$ and $\nu_{\text{nom}} = (45, 1, 10, 10)$. The main difference between the coupled and decoupled implementations of the three VPP control policies is in the available downward/negative flexibility. That is, ESSs can be used to control down flexibility in the decoupled case while EWHs and EV can only control flexibility upwards. The coupled case does not allow for such behavior since the rate of acceptance of charging packets is the same for all DER types, which limits the down flexibility provided by ESSs.

TABLE III
RMS TRACKING ERROR AND ONE-STEP-AHEAD UP/DOWN FLEXIBILITY

Tracking	RMS (kW)			$\overline{flex}$ (kW)			$\underline{flex}$ (kW)		
	Max	Min	QoS	Max	Min	QoS	Max	Min	QoS
Coupled	24.8	13.19	15.73	166.6	383.7	288.6	196.5	424.5	320
Decoupled	28.7	19.6	18.3	161.4	633.1.0	342.5	200.8	357.6	325.7

TABLE IV
QoS RMS SET-POINT DEVIATION FOR EWHs (°C),
ESSs (%) AND PEVs (%)

RMS	Max			Min			QoS		
	EWH	ESS	PEV	EWH	ESS	PEV	EWH	ESS	PEV
Coupled	1.05	7.65	4.18	0.39	1.81	1.45	0.54	1.22	0.79
Decoupled	1.47	7.52	0.93	0.74	1.08	3.74	0.19	0.71	0.56

TABLE V
POPULATION AVG. ST. DEV. EWHs (°C),
ESSs (%) AND PEVs (%)

QoS Std	Max			Min			QoS		
	EWH	ESS	PEV	EWH	ESS	PEV	EWH	ESS	PEV
Coupled	1.11	8.99	14.3	1.2	8.27	14.9	1.2	8.57	14.8
Decoupled	1.01	8.91	14.9	1.24	8.34	14.38	1.25	8.42	14.82

2) QoS Performance: One can also observe in these figures and in Table IV that the average dynamic state of each DER class tends to be around the desired set point, which ensures that the average DER satisfies QoS requirements while the population is tracking the reference. Specifically, Table IV shows the RMS tracking error values of the average QoS with respect to the desired set point.

However, the average dynamic state could be misleading if the DER populations show too much variation over the dead-band or beyond it. In this regard, Table V presents the average standard deviation for each DER population under the control laws described in Theorem 1. One can infer from these values that while TCLs and ESSs populations are relatively close to their average dynamic state (≈ 1°C and around 9% for ESSs), the EV population have an almost constant standard deviation of around 14.5%, which is a result of EVs uncontrollably transitioning to driving (i.e., discharging) events. Note that future work will focus on control mechanisms for allowing PEM to not only regulate with respect to the average population QoS set-point but to actively reduce the variance in the population (beyond opt-out control), while also tracking the reference. Table V illustrates that managing a fleet relative to the QoS set-points does not guarantee an overall reduction in variance.

Note that when the reference signal is centered around the nominal power level of the system, then QoS set-points can be attained. Furthermore, as mentioned in Section III-B, due to the presence of multiple DER types, several combinations
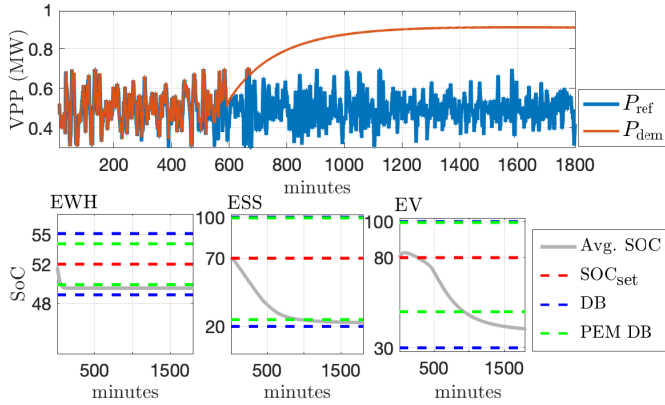
Fig. 11. Fleet of 1000 TCLs, 1000 ESSs, and 250 EVs tracking a signal below nominal power. Top plot: reference tracking by minimizing number of request. Bottom plots: SOC for each individual macromodel showing that the SOCs reach critical levels and, therefore, the populations start opting out as designed by the PEM scheme.

of charging and discharging requests can be accepted so that the tracking error is zero. However, the QoS objective of (10) determines the number of requests to accept that ensures QoS as well as the control input β is close to the nominal obtained from (7). In the absence of $\left|\chi_{c,d}[k] - \chi_{c,d}^{\text{nom}}\right|$ term in (10), the VPP has no information about the nominal control input and hence determines $\chi_{c,d}$ that minimizes QoS deviation from the set-point at the next time-step depending only on the current SOC of each population. This results in prioritizing one class of DERs whose SOC is farthest from the set-point at the current time step. Moreover, the $\chi_{c,d}$ obtained without the nominal information is not guaranteed to minimize QoS deviation over longer time periods. Recall that in PEM, the DERs whose packet requests have been accepted, charge or discharge for a time equal to at least the packet length. As a result, oscillations are produced in the power consumption of individual DER population while the aggregate still tracks the provided reference with minimal tracking error. One can potentially avoid such oscillations by optimizing QoS over a prediction horizon in a model predictive control (MPC) formulation, similar to the one in [28]. However, this is not trivial due to the bin based population model, studied in this work, being nonconvex and having a large number of states. Further work is required to develop a low-order model that admits an MPC type formulation and is the topic of ongoing research. Nevertheless, including the nominal information along with QoS ensures that the VPP obtains $\chi_{c,d}$ that satisfies QoS while the control input is close to the desired nominal values.

Finally, Fig. 11 shows the tracking of a signal that is well below the nominal power and, therefore, the system starts losing energy as shown in the SOC plots. Clearly, DERs start opting out as designed by the PEM scheme around minute 350. Different DER classes opt-out at different rates. The figure also shows that as the overall SOC of the system saturates, the system is unable to keep tracking the signal. Again, to be clear, the SOC of the EV population is decreasing because insufficient charging requests are accepted by the VPP compared with the discharging from end-use driving events.

IV. PARAMETRIC HETEROGENEITY OF DERs IN PEM

Thus far, a diverse VPP has been defined as one that coordinates different classes of DERs (e.g., EWH, ESS, and PEV) via coupled or decoupled macromodels as shown in Fig. 5. However, a single macromodel, as described in Section II-B, represents a population of homogeneous DERs that are all of the same type (e.g., EWHs with the same parameters). Furthermore, it is valuable to be able to employ the macromodel to accurately represent intra-DER-class heterogeneity. The goal of this section is to analyze and quantify the tracking error when a parametrically heterogeneous population of one specific class of DERs (with respect to their rated power parameter) is approximated by a group of homogeneous macromodels. The reason to focus on the rated power parameter is due to PEM's unique request-response mechanism for coordination that uses an estimate of rated power to compute the control action, e.g., the computation of β_c and β_d in Section III requires knowledge of P_c^{rate} and P_d^{rate}. This introduces uncertainty in the coordination of an intra-DER-class heterogeneous population. Hence, the tracking error is due to the fact that control inputs are computed with respect to a finite discrete group rather than the continuous underlying distribution that defines the heterogeneous parameter variations.

Consider a population of residential EWHs from different manufacturers whose temperatures evolve as

$$z_n[k+1] = z_n[k] + \Delta t \left(\frac{\eta_n u_{\phi_n,n}}{c_p \rho L_n} - \frac{z_n[k] - z_a}{\tau_n} - \frac{z_n[k] - z_{\text{in}}}{60 L_n} w_n[k] \right) \quad (11)$$

where z_n is the average tank temperature [°C] corresponding to DER n, c_p is the specific heat constant, ρ is the density of water when close to $50\,°\text{C}$ and L_n is liters in the container tank. Furthermore, $z_{a,n} \equiv z_a$ is the ambient temperature, $\tau_n \equiv \tau$ is the standing loss time constant to ambient temperature, $\eta_n \equiv \eta$ for all EWHs and w_n is the end-use process as described in Section II-A. The parameters are the same as in [19, Table 1]. The mechanical input $u_{\phi_n,n} = P_{c,n}^{\text{rate}}$ [kW], if logic state $\phi_n = c$ (charging) and $u_{\phi_n,n} = 0$, if $\phi_n = s$ (standby). Note that (11) is a particular case of (1). In addition, the model in (11) is heterogeneous only in $P_{c,n}^{\text{rate}}$. Each EWH is simulated individually according to the request-response mechanism of PEM described in Section II and is referred to as *agent-based* simulations. The rated power heterogeneity considered in this section is introduced next.

The rated power for the nth EWH is assumed to be uniformly distributed over some fixed power rating interval. This forms a *rated-power-heterogeneous* population of packetized EWHs. Each EWHs is simulated in an agent-based simulation environment and the population is divided into groups based on their rated power operating under the same VPP and denoted as group-based simulations. The VPP receives measurements of the total power demand of the population and determines the number of requests to be accepted. This decision depends upon the information available to the VPP regarding the groups. The objective, therefore, is to compare the behavior of a controlled rated-power-heterogeneous population of packetized DERs divided into g groups, with that

of an aggregation of g homogeneous macromodels. The rated power associated with the ith macromodel then corresponds to the average rated power of that group. The set of g intra-DER-class groups under PEM are in an arrangement similar to that in Fig. 5.

To illustrate the idea, let a rated-power-heterogeneous population over the power interval [4, 12] kW be used to track a variable regulation signal. An increasing finite set of evenly separated values $\mathcal{P} := \{\bar{P}_1, \ldots, \bar{P}_g\}$ with $\bar{P}_j \in [4, 12]$ kW is used to group the devices. That is, an EWH with rated power P_i belongs to the jth group if $\bar{P}_j$ makes $|P_i - \bar{P}_j|$ be the smallest with respect to all other values in $\mathcal{P}$. Fig. 12 shows the tracking results for a rated-power-heterogeneous population of EWHs compared against $1, 2, 4$ groups having rated power evenly distributed over [4, 12] kW while tracking the same regulation signal. Note here that $g = 1$ indicates a group characterized by rated power of 8kW while $g = 4$ implies groups with rated powers $\{5, 7, 9, 11\}$ kW. The key takeaway is that the VPP makes decisions with respect to the rated power value $\bar{P}_j$ of the groups whereas, the rated power of EWHs is uniformly distributed over the mentioned interval. It is also important to highlight that for EWHs there is no difference between the control strategies in Section III since there are no discharging packets available.

As expected, the tracking error is lower when more groups are used. In Fig. 12, the VPP assumes that all requests from the ith group has rated power $\bar{P}_i$ instead of the exact value which is uniformly distributed. Clearly, if an accurate day-ahead or multi-hour prediction of flexibility from a VPP was critical for feed-forward control (e.g., optimizing VPP's response to a peak demand event), then $g > 2$ groups are helpful. However, if an observer is implemented to provide an updated estimate of the states, say, every minute, then it is foreseeable that one or two groups could suffice for control. Estimation and prediction of QoS is an important topic and represents ongoing work that is outside the scope of this manuscript. For the sake of simplicity, the analysis hereafter is limited to EWHs, therefore, only charging requests are considered. Furthermore, in Fig. 12, the VPP responds sequentially to packet requests whereas in the macromodel, the VPP gathers all requests within a time interval and determines the total number of requests to accept as discussed next.

A. Comparing Synchronous Model With Asynchronous Reality

By design, PEM is asynchronous and, therefore, responds to each energy packet request in a *sequential* manner. This means that the controller keeps accepting requests while demand power is less than the reference. This is similar to event-driven, real-time, cloud-based but serverless implementations of large-scale IoT data processing applications. A request in that scenario is an event that triggers a response from the load coordinator amounting to accept or reject the packet request [41]. On the other hand, the macromodel from Section II-B is implemented synchronously and, therefore, *gathers* all requests over a time step interval, say $[k\Delta t, (k + 1)\Delta t]$, and then computes the proportion of packet requests
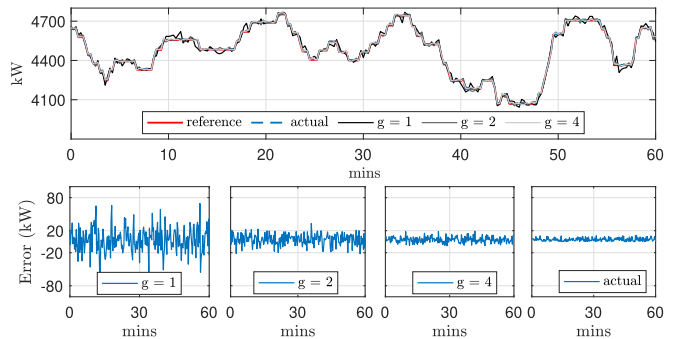

Fig. 12. (Top) Reference-tracking comparison between a VPP agent-based (micro-)model with 10 000 heterogeneous EWHs with rated power uniformly distributed over [4, 12]kW and g groups each with $10\,000/g$ EWHs evenly divided into groups of constant rated power. (Bottom) The tracking error for different groupings of the 10 000 EWHs.

to accept, β, using the guidelines in Section III. This section establishes a relationship between the sequential and gathered implementations of PEM in the presence of parametric heterogeneity as modeled by g homogeneous groups.

First, consider a homogeneous population of DERs and compare sequential and gathered implementations of PEM. In the synchronous (gathered) case, the PEM coordinator receives requests as a queue. More precisely, let n_t be the number of requests received over a time step interval and $n_a \le n_t$ be the number of accepted requests. The sequence of requests is denoted by $s_r := (r_i)_{i \le n_t}$, r_i is the ith request received and P_i is the associated rated power of such request. Denoting by $\mathcal{S}_{n_t}$ the group of all bijections from the set of n_t elements on itself, a permutation Θ is an element of $\mathcal{S}_{n_t}$ from a given set of elements. An element $\Theta \in \mathcal{S}_{n_a}$ is written as

$$\Theta = \begin{pmatrix} 1 & 2 & \cdots & n_a \\ \theta(1) & \theta(2) & \cdots & \theta(n_a) \end{pmatrix}.$$

If the permutation is realized in a random fashion, then it is called a *random permutation*. It then follows that another realization of the request process during a time interval produces a random permutation $\Theta(s_r)$. If, in addition, the population is homogeneous with the same rated power for each request, then $\Theta(s_r)$ is indistinguishable of s_r because the output power they contribute is the same. Now, given a sequence s_r, asynchronous (sequential) PEM accepts the first n_a requests based on order of arrivals. On the other hand, gathered PEM accepts request based on the entire (unordered) sequence s_r by computing the percentage of n_t that balances demand and the reference. Note in this case that there are $\binom{n_t}{n_a}$ sequences that can be chosen by the coordinator from the n_t received requests. In fact, all these sequences form an equivalence class with respect to their output power. Moreover, the output power from this equivalence class is exactly the same as that of the sequential PEM acceptance methodology. This fact indicates that the macromodel for a homogeneous population of DERs can accurately represent PEM.

The case is more complex for heterogeneous populations approximated by g groups of homogeneous DERs. Under sequential PEM, incoming requests may have different rated powers, one can think of g homogeneous sequences $s_{r,i}$ for $i = 1, \ldots, g$. A request received at any particular time

can come from any of the g groups depending upon their probability of request. To model such a string of request, let $n_{t,i}$ be the total number of request in group i and r^i_j be the jth request of the ith group at time $\tau^i_j \in \mathbb{R}$. The set $\tau^i := \{\tau^i_1, \ldots, \tau^i_{n_{t,i}}\}$ is naturally ordered since $\tau^i_j < \tau^i_{j+1}$. For a set $A = \{a_1, a_2, \ldots, a_n\}$ with $a_i \in \mathbb{R}$, the ordering operation $\mathrm{ord}_<(.) \in \mathcal{S}_n$ permutes A into $\mathrm{ord}_<(A) = \{a_{\mathrm{ord}_<(1)}, a_{\mathrm{ord}_<(2)}, \ldots, a_{\mathrm{ord}_<(n)}\}$ such that $a_{\mathrm{ord}_<(i)} < a_{\mathrm{ord}_<(i+1)}$. Defining $\bar{\tau} = \mathrm{ord}_<(\tau^1 \cup \cdots \cup \tau^g)$, the sequence $s_r = (r_i)_{i \leq n_t}$ such that $r_i{}^2$ is the request associated to $\bar{\tau}_i \in \bar{\tau}$ constitutes the string of request received sequentially by the coordinator. The coordinator now accepts the first incoming n_a requests from s_r in similar fashion to the homogeneous case. For the gathered mechanism when $g > 1$, the acceptance rate is computed as in the coupled case of diverse populations in Section IV. That is, one computes the overall accepting rate as

$$\beta = \frac{n_a}{n_t} \tag{12}$$

where n_a is as before the total number of accepted requests gathered from the g groups without any specific order of arrival. Then β is applied equally to all the groups.

The objective is now to relate the sequential and gathered mechanisms when $g > 1$. To achieve that, first observe that the power output of the accepted request now depends specifically on how the requests from different groups arrive and from the sequence of accepted requests of length n_a. For example, the first n_a requests could come from the population with minimum rated power, P_1, or from the population with maximum rated power, P_g. Clearly, $n_a P_1 \neq n_a P_g$. Nevertheless, the distribution characterizing all the possible sequences of length n_a drawn from the n_t requests from the g groups obey a multivariate hypergeometric distribution (MHD). Considering that requests from the same group are indistinguishable, the MHD gives the likelihood of a particular sequence to appear in a realization of the requesting process. Let $n_{a,i}$ be the number of accepted requests in group i. The acceptance rate for the ith group is $\beta_{\mathrm{avg},i} = n_{a,i}/n_i$, where n_i denotes the total number of requests received from group i (with mild abuse of notation). From the properties of the MHD [42], the expected number of accepted requests per group is $E[n_{a,i}] = n_a n_i/n_t$. The average of the expected accepting rates is then

$$\beta_{\mathrm{avg}} = \frac{1}{g} \sum_{i=1}^{g} \frac{E[n_{a,i}]}{n_i} = \frac{1}{g} \sum_{i=1}^{g} \frac{n_a}{n_t} = \frac{n_a}{n_t}$$

which is the same as (12). In fact, for every i, one has that $E[\beta_{\mathrm{avg},i}] = E[n_{a,i}]/n_i = n_a/n_t$, which is constant and independent of i. Hence, the expected packet acceptance rate over the set of all possible sequences of length n_a obtained from the received request sequence s_r with g type of requests is equivalent to the even application of (12) to all the groups. This result will allow the computation of bounds on tracking error using the PEM macromodel and the control schemes presented in Section III.

---

² r_i can come from any of the s_{r_i} sequences forming s_r.

B. Group-Based Tracking Error and Information Scenarios

The objective now is to quantify the VPP tracking error by considering each group being modeled by a macromodel that has an estimate of the group's DERs' rated powers. Depending on the distribution of the rated powers (P_i), the macromodel groups can differ in their population count (N_i). As discussed in Section III, the VPP accepts/denies energy packet requests based on the tracking error (or gap)

$$P_{\mathrm{gap}} := P_{\mathrm{ref}} - P_{\mathrm{dem}} + \sum_{i=1}^{g} P_i n^i_{\mathrm{sb},c} \tag{13}$$

where P_{dem} represents, as before, the measured power demand of the entire DER population, $n^i_{\mathrm{sb},c}$ is the total number of packets expiring from i-th group, and $\sum_{i=1}^{g} P_i n^i_{\mathrm{sb},c}$ is the power that will be lost due to $n^t_{\mathrm{sb},c} := \sum_{i=1}^{g} n^i_{\mathrm{sb},c}$ packets expiring since only EWHs are considered and there are no discharging packets. The information regarding the total number of expiring packets is always available to the VPP from the decision making process and from the timer states, as described in Section II-B, which act as ledger for the expiration of packets.

Recalling that the current total number of packet requests from the ith group is n_i and that the total number of requests from g groups is $n_t = \sum_{i=1}^{g} n_i$. In a similar manner, the total number of EWHs in the entire population is $N_t = \sum_{i=1}^{g} N_i$, where N_i is the number of DERs in the ith group. Without loss of generality, assume $P_1 < P_2 < \cdots < P_g$, where P_i is the rated power used by the ith group. Under this setting, the VPP coordinator receives a total n_t number of requests from the packetized EWHs and the total power demand P_{dem}. However, the VPP does not know the power associated with each of these requests. Therefore the rated power of the requesting population P is estimated by the VPP in order to compute a scalar control input β (i.e., the proportion of requests to accept)

$$\beta = \begin{cases} 1, & \text{if } P_{\mathrm{gap}} > P n_t \\ 0, & \text{if } P_{\mathrm{gap}} < 0 \\ \frac{P_{\mathrm{gap}}}{P n_t}, & \text{otherwise.} \end{cases} \tag{14}$$

The focus hereafter is when $0 < P_{\mathrm{gap}} \leq P n_t$. From Section III, the VPP's control objective tracks the reference to minimize P_{gap}. Thus, the VPP accepts βn_t requests and packetized demand increases by

$$P_{\mathrm{acc}} := \beta \left(\sum_{i=1}^{g} P_i n_i \right). \tag{15}$$

Note in (14) that the coordinator's estimate of the requesting power rating P is used explicitly to reduce P_{gap} (i.e., track the reference) since individual rated powers are not communicated to the VPP. Thus, from (14) and (15), the values N_i, P_i and n_i affect the accuracy in which devices supply power in order to make P_{gap} as small as possible. Defining tracking error $P_{\mathrm{err}} = P_{\mathrm{gap}} - P_{\mathrm{acc}}$, then it follows that

$$P_{\mathrm{err}} = \left(1 - \frac{\sum_{i=1}^{g} P_i n_i}{P n_t} \right) P_{\mathrm{gap}}. \tag{16}$$

It should be noted here that the VPP assumes the packet size of all requests to be P which is an estimate depending upon the information available to the VPP. In this regard, three information scenarios are considered as illustrated in Fig. 13. Furthermore, analytical bounds on tracking error in each of these scenarios are explicitly derived and validated next.

1) *Full Information:* It is available at all times: i.e., P_i, N_i, and n_i are known for all the groups. Hence, the VPP's estimate of P is computed as the weighted average $P^* = \sum_{i=1}^{g} P_i n_i / n_t$. Since only packets of fixed power are allowed, the main source of mismatch between accepted packets and P_{gap} is roundoff error. The roundoff errors is considered coming from two cases: sequential and gathered implementations of PEM. Since, the VPP makes decisions with full knowledge of the power of each request, the error incurred P_{err}^* in sequential PEM is bounded by

$$\left| P_{\text{err}}^{*,1} \right| \leq \frac{P_g}{2}. \tag{17}$$

Relation (17) is due to the fact that the last accepted request brings P_{gap} within half a packet. However, in the gathered PEM case, the proportion of accepted requests is applied equally to all the groups, therefore

$$\left| P_{\text{err}}^{*,2} \right| \leq \sum_{i=1}^{g} \frac{P_i}{2}. \tag{18}$$

Equation (18) tells that each group incur into a roundoff error of half the size of its corresponding P_i. Hence, gathered PEM with groups is not implemented in the field since it amplifies the roundoff error.

2) *Static Information:* It is available offline: only N_i and P_i are known for all the groups. The resulting estimate of rated power for the entire population is then

$$P' := \frac{\sum_{i=1}^{g} P_i N_i}{N_t}. \tag{19}$$

In this scenario, the coordinator is unaware of the rated power of individual requests, therefore, it estimates P_{gap} from the total number of requests n_t and the total number of expiring packets $n_{\text{sb},c}^t$ as

$$P_{\text{gap}}'[k] := P_{\text{ref}}[k] - P_{\text{dem}}[k] + P' n_{\text{sb},c}^t \tag{20}$$

where $P' n_{\text{sb},c}^t$ and $P' n^t$ are the coordinator's best estimate of the demand reduction due to packets expiring and total power associated with packet requests respectively, in the time-step k. The control β' is then computed from P_{gap}' and P' as

$$\beta' = \frac{P_{\text{gap}}'}{P' n_t} \tag{21}$$

where $\beta' = 0$ if $P_{\text{gap}}' < 0$ and $\beta' = 1$ when $P_{\text{gap}}' > P' n_t$. The increase in demand after accepting $\beta' n_t$ requests is

$$P_{\text{acc}}' := P' \beta' n_t = P_{\text{gap}}'. \tag{22}$$

Note that (22) holds only when $P_{\text{gap}}' \in [0, P' n_t]$ since (21) was used and $\beta' \in [0, 1]$. Recall also that there
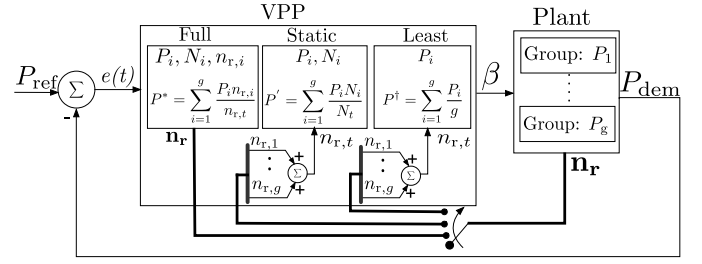

Fig. 13. PEM system for heterogeneous population of the same type of DER. Known parameters at the VPP in *Full, Static* and *Least* information cases is included in corresponding boxes.

is an roundoff error due to PEM as well as the error introduced due to difference realizations of the request sequence presented in Section IV-A. Thus, using (13), (17) and (20), the VPP tracking error is

$$P_{\text{err}}' = P_{\text{err}}^{*,1} + P_{\text{gap}} - P_{\text{gap}}'$$
$$= P_{\text{err}}^{*,1} + \sum_{i=1}^{g} P_i n_{\text{sb},c}^i - n_{\text{sb},c}^i P' \pm P_{\text{err},\Theta}' \tag{23}$$

where, $P_{\text{err},\Theta}'$ is the error due to different realizations of request sequence and is explicitly derived in the next paragraph.

3) *Least Information:* It is available offline: only P_i is known for every group. Specifically, the VPP has no information about requests from each group (n_i) as well as the composition of each group (N_i). Then a pragmatic estimate of the rated power for the entire population is the simple average, i.e., $P^\dagger = 1/g \sum_{i=1}^{g} P_i$. The VPP's estimate of $P_{\text{gap}}^\dagger$ is obtained by replacing P' with $P^\dagger$ in (20). The corresponding tracking error $P_{\text{err}}^\dagger$ is similar to (23) but with P' replaced by $P^\dagger$ and the corresponding $P_{\text{err},\Theta}^\dagger$ instead of $P_{\text{err},\Theta}'$ which is derived next.

In addition, one can provide conservative bounds for tracking errors P_{err}^*, P_{err}', and $P_{\text{err}}^\dagger$. With the availability of *Full information*, the only tracking error is due to roundoff error. The error bounds follow directly from (17) as

$$P_{\text{err,min}}^* := -\frac{P_g}{2} \leq P_{\text{err}}^{*,1} \leq \frac{P_g}{2} =: P_{\text{err,max}}^*. \tag{24}$$

The bounds on tracking error in *Static* and *Least* information cases are obtained by quantifying $P_{\text{err},\Theta}'$ and $P_{\text{err},\Theta}^\dagger$, respectively, in (23). Consider first the *Static* information case where P' and P_{err}' are the inferred rated power by the coordinator and the corresponding error caused by such a choice respectively. *Least* information case will then follow from the same procedure. The error due to the assumption that expiring packets are of the same rated power P' is characterized by (23). However, as described in Section IV-A, different realizations of the request sequence provide different output powers to balance P_{gap}. The statistics of the MHD is used to bound the error due to realizations of the request sequence s_r, denoted as $P_{\text{err},\Theta}'$ in (23). The difference is now that the number of requests per group n_i are unknown. But one can still rely on the fact that N_i is known in the *Static information* case. That is, under the assumption that the g groups have similar

probabilities of request (see Section II-A), the estimated value of n_i is $\hat{n}_i = n_t N_i / N_t$. The expectation value of $n_{a,i}$ is

$$E[n_{a,i}] = \frac{n_a \hat{n}_i}{n_t} = \frac{n_a N_i}{N_t}$$

and its variance is

$$\begin{aligned} \text{Var}[n_{a,i}] &= \frac{\hat{n}_i}{n_t}\left(1 - \frac{\hat{n}_i}{n_t}\right)n_a\left(\frac{n_t - n_a}{n_t - 1}\right) \\ &= \frac{N_i}{N_t}\left(1 - \frac{N_i}{N_t}\right)n_a\left(\frac{n_t - n_a}{n_t - 1}\right) \end{aligned} \quad (25)$$

where (25) follows directly from the MHD [42]. Furthermore, observe in (25) that all quantities are known except for n_a. A straightforward maximization of (25) gives

$$n_{a,\max} := \frac{n_t}{2}.$$

Note that $n_{a,\max}$ is independent of i, therefore one has that (25) is upper bounded by $\sigma^2 := \text{Var}[n_{a,i}]|_{n_a = n_{a,\max}}$. Furthermore, defining $\beta_\sigma = g\sqrt{\sigma}/n_t$, it follows that the power accepted is redefined as

$$P'_{\text{acc}} := P'(\beta' \pm \beta_\sigma)n_t = P'_{\text{gap}} \pm P'_{\text{err},\Theta} \quad (26)$$

where $P'_{\text{err},\Theta} := P'g\sqrt{\sigma}$ and contains the error due to the variations in request acceptance from each group. Using (21) and (26) one can obtain (23) after including roundoff error.

As an example, consider the *Static* information case with $g = 2$, $N_t = 2500$, $N_1 = 2000$, $N_2 = 500$, and $n_t = 20$. Then, $n_{a,\max} = 10$, $\sigma = 0.84$ and $\beta_\sigma = 0.0918$. The latter additional acceptance rate amounts to $P'_{\text{err},\Theta} = 1.8\ P'$ extra error in power accepted. In general, one can include $K\beta_\sigma$ with $K \in \mathbb{N}$ in (26) instead of just β_σ depending on how much of the distribution tail of $n_{a,i}$ needs to be captured. It is shown by the simulations in the next section that $K = 1$ suffice in capturing the tracking errors described previously.

The tracking error bounds for the *Static information* case are now obtained as follows: P'_{err} in (23) achieves its minimum ($P'_{\text{err},\min}$) when the total number of expiring packets $n^i_{\text{sb},c}$ belong to group with the smallest rated power (P_1). This occurs when $n^i_{\text{sb},c} = 0$ for $i \neq 1$ (i.e., $n^t_{\text{sb},c} = n^1_{\text{sb},c}$). Similarly, P'_{err} is maximized ($P'_{\text{err},\max}$) when $n^i_{\text{sb},c} = 0$ for $i \neq g$ (i.e., $n^t_{\text{sb},c} = n^g_{\text{sb},c}$). The maximum and minimum bounds are

$$P'_{\text{err},\max} := P^*_{\text{err},\max} + (P_g - P')n^t_{\text{sb},c} + P'_{\text{err},\Theta}, \quad (27)$$
$$P'_{\text{err},\min} := P^*_{\text{err},\min} + (P_1 - P')n^t_{\text{sb},c} - P'_{\text{err},\Theta}. \quad (28)$$

For the case of *Least information*, the bounds on tracking error are obtained by replacing P' with $P^\dagger$ in (26) through (28) and $P^\dagger_{\text{err},\Theta} := P^\dagger g\sqrt{\sigma}$. Particularly for the *least information* case, one has assumed that $N_i = N_t/g$ since there is no other offline information available. The next section validates the tracking error bounds for the three information scenarios.

C. Simulations

Consider a heterogeneous population consisting of 4000 EWHs with rated power 6 kW and 2000 EWHs with rated power 10 kW. Fig. 14 shows the results of the agent based simulations of the heterogeneous population tracking a scaled and shifted reference signal. In Fig. 14, it can be seen that
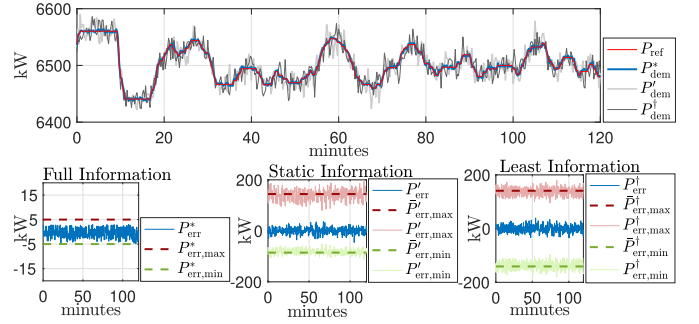

Fig. 14. (Top) Total power demand of two groups and (Bottom) tracking error for the *Full*, *Static* and *Least information* cases.

TABLE VI
TRACKING ERROR FOR TWO GROUPS $(P_1, P_2) = (6, 10)$
AND $(N_1, N_2) = (4000, 2000)$

Error	P_{err}			$P_{\text{err, max}}$			$P_{\text{err, min}}$		
	Max	RMS	Min	Max	RMS	Min	Max	RMS	Min
Full	2.98	1.96	-4.95	-	-	-	-	-	-
Static	36	12.3	-45.5	201.5	144.3	82.8	-54.7	-85.4	-114.4
Least	39.6	12.9	-34.7	182	140.8	98.4	-98.4	-140.8	-182

the *full information* case (with P^*), results in the tracking error $|P^*_{\text{err}}| \leq P_2 = 5$ kW according to (24) and the RMS tracking error is 1.96 kW. The *Static information* case produces tracking error bounds according to $P'_{\text{err},\max}$ into (27) and (28). The tracking error is always within $P'_{\text{err},\max}$ and $P'_{\text{err},\min}$ and the RMS tracking error is 12.3 kW, whereas the maximum and minimum tracking error observed is 36 and -45.5 kW respectively. These bounds are a function of the number of expiring packets and hence a degree of variation is observed in Fig. 14. The RMS values of the maximum and minimum bounds ($\bar{P}'_{\text{err, max}}$ and $\bar{P}'_{\text{err, min}}$, respectively) are also plotted with dashed lines in Fig. 14. The RMS tracking error for the *Least information* case is 12.9 kW and the maximum and minimum values of the tracking error are 39.6 and -34.7 kW respectively. The mean error due to the *Least information* case's estimate is larger than the one for the *Static information* case, but they are not expected to disagree by much since $P' = 8$ kW and $P^\dagger = 7.3$ kW. Furthermore, the dashed lines in Fig. 14 labeled as $\bar{P}^\dagger_{\text{err, max}}$ and $\bar{P}^\dagger_{\text{err, min}}$ are the RMS values of the maximum and minimum error bounds respectively. Table VI summarizes the experiment.

The second simulation aims to illustrate the effect on tracking when populations vary significantly in rated powers. Consider two groups $(P_1, P_2) = (3, 12)$ kW with $(N_1, N_2) = (2000, 500)$ EWHs. The results are shown in Fig. 15 and the corresponding errors are tabulated in Table VII. The RMS tracking error in the presence of *Full information* is 1.64 kW which is within ± 6 kW according to (24). For the *Static information* case, the coordinator's inferred power rating of the population is $P' = 4.8$ kW which is biased towards the group with rated power P_1. The RMS tracking error is 17.3 kW and the maximum and minimum observed errors are 64.5 and -55.5 kW, respectively. Similarly, for the *Least information* case, the RMS tracking error is 19.5 kW

TABLE VII
TRACKING ERROR FOR TWO GROUPS $(P_1, P_2) = (3, 12)$ kW
AND $(N_1, N_2) = (2000, 500)$

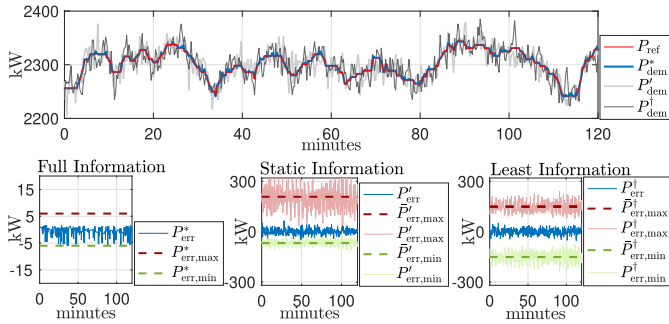| Error	$\mathbf{P_{err}}$			$\mathbf{P_{err, max}}$			$\mathbf{P_{err, min}}$		
	Max	RMS	Min	Max	RMS	Min	Max	RMS	Min
Full	1.5	1.64	-5.86	-	-	-	-	-	-
Static	64.5	17.3	-55.5	361.6	208	37.1	-26.3	-68.132	-107.7
Least	65.2	19.5	-56.8	228.3	151.3	65.8	-65.8	-151.3	-228.3

Fig. 15.   (Top) Total power demand of two groups and (Bottom) tracking error for *Full*, *Static* and *Least information* cases.

along with the maximum and minimum errors of 65.2 and -56.8 kW, respectively. Furthermore, it should be noted that the error bounds in the *Static information* case are tighter compared to the *Least information* case. The reason is that the coordinator with *Least information* infers power rating $P^\dagger = 7.5$ kW and ignores the size of each group. Furthermore, it should be noted here that although $P^\dagger$ is less accurate as compared to the *Static* information case, however, the worst case bounds account for such inaccuracies by considering all expiring packets either belong to the group with rated power 3 kW ($P^\dagger_{err,min}$) or to the group consisting of 12-kW EWHs ($P^\dagger_{err,max}$). Finally, as in Fig. 14, the dashed lines in Fig. 15 are the RMS values of the maximum and minimum bounds on tracking error.

As expected, the *Full* or *Static information* cases outperform the *Least information* case, which highlights the performance gains possible in PEM by including static and dynamic information in each request sent to the VPP. Of course, more information per packet request requires larger data exchanges and leads to larger communication overhead for PEM, which represents a possible trade-off between communication and control. This topic represents future work.

V. CONCLUSION

Reference-tracking control policies were analyzed and presented for aggregated, diverse DER populations operating under PEM. This allowed for real-time coordination of VPP resources in a power balancing scenario. Specific to PEM, a QoS-preserving nominal response was defined and used to compute an effective power reference signal that enables effective and sustained tracking. Several scenarios were presented to validate the control schemes. Among these, were different cases of information sharing among DER groups of packetized DERs. In addition, control policies were presented

that optimize (maximize/minimize) the number of accepted packet request and prioritize QoS with respect to a desired set point for the DERs. The latter case highlights interesting, undesirable effects arising from the DER population dynamics constrained to maintain specific QoS levels and requires prioritization of flexible DER classes. Performance metrics developed include RMS tracking error, up/down power flexibility, and average standard deviation of the DER populations. Additionally, a grouping strategy for handling intra-class DER heterogeneity under PEM was developed. First, sequential and gathered accepting mechanisms were studied and related to each other showing that gathered acceptance constitutes the averaging scenario for all possibilities of sequential accepting. Finally, bounds on reference tracking performance were quantified analytically and validated with simulations to describe the effects of approximating intra-class heterogeneous populations of DERs (relative to their rated power) with groups of homogeneous macromodels.

Future work will consider including m_R and packet size δ in the optimization problem (7) to obtain the nominal control $\beta^*(m_R^*, \delta^*)$. This will allow the VPP to select the best probability of request curve corresponding to each DER class. Furthermore, the analysis of the error due to the aggregation processes as in [33] for the case of EWHs will improve the quantification of the flexibility provided by PEM. One can then expand the heterogeneous assumption to include thermal parameters (e.g., tank capacity and insulation). Finally, state-estimation and predictive control policies for PEM based on the average *state of charge* in the form of observer design and a low-order *virtual battery* model [38] will also be investigated. The state estimator will build on preliminary results in [18] where a Kalman filter performed well for the macro-model under some simplifying assumptions.

DISCLOSURE

REFERENCES

[1] N. J. Cammardella, R. W. Moye, Y. Chen, and S. P. Meyn, "An energy storage cost comparison: Li-ion batteries vs distributed load control," in *Proc. Clemson Univ. Power Syst. Conf. (PSC)*, Sep. 2018, pp. 1–6.

[2] K. Desrochers *et al.*, "Real-world, full-scale validation of power balancing services from packetized virtual batteries," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Feb. 2019, pp. 1–5.

[3] H. Lo, S. Blumsack, P. Hines, and S. Meyn, "Electricity rates for the zero marginal cost grid," *Electr. J.*, vol. 32, no. 3, pp. 39–43, Apr. 2019.

[4] M. G. Morgan and S. N. Talukdar, "Electric power load management: Some technical, economic, regulatory and social issues," *Proc. IEEE*, vol. 67, no. 2, pp. 241–312, Feb. 1979.

[5] 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 App. Syst.*, no. 3, pp. 1151–1163, May 1980.

[6] A. Brooks, E. Lu, D. Reicher, C. Spirakis, and B. Weihl, "Demand dispatch," *IEEE Power Energy Mag.*, vol. 8, no. 3, pp. 20–29, May/Jun. 2010.

[7] D. S. Callaway and I. A. Hiskens, "Achieving controllability of electric loads," *Proc. IEEE*, vol. 99, no. 1, pp. 184–199, Jan. 2011.

[8] K. M. Brokish, "Adaptive load control of microgrids with non-dispatchable generation," M.S. thesis, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.

[9] J. Frolik and P. Hines, "Random access, electric vehicle charge management," in *Proc. IEEE Int. Electric Vehicle Conf.*, Mar. 2012, pp. 1–4.

[10] J. L. Mathieu, S. Koch, and D. S. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 430–440, Feb. 2013.

[11] S. P. Meyn, P. Barooah, A. Busic, Y. Chen, and J. Ehren, "Ancillary service to the grid using intelligent deferrable loads," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2847–2862, Nov. 2015.

[12] J. Frolik and P. Hines, "Systems and methods for random-access power management using packetization," U.S. Patent 10 256 631 B2, Apr. 9, 2015.

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

[14] M. Almassalkhi, J. Frolik, and P. Hines, "Packetized energy management: Asynchronous and anonymous coordination of thermostatically controlled loads," in *Proc. Amer. Control Conf. (ACC)*, May 2017, pp. 1431–1437.

[15] M. Almassalkhi, L. A. Duffaut Espinosa, P. D. Hines, J. Frolik, S. Paudyal, and M. Amini, *Asynchronous Coordination of Distributed Energy Resources With Packetized Energy Management*. New York, NY, USA: Springer, 2018, pp. 333–361.

[16] L. A. Duffaut Espinosa, M. Almassalkhi, P. Hines, S. Heydari, and J. Frolik, "Towards a macromodel for packetized energy management of resistive water heaters," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2017, pp. 1–6.

[17] L. A. Duffaut Espinosa, M. Almassalkhi, P. Hines, and J. Frolik, "Aggregate modeling and coordination of diverse energy resources under packetized energy management," in *Proc. IEEE 56th Annu. Conf. Decis. Control (CDC)*, Dec. 2017, pp. 1394–1400.

[18] L. A. Duffaut Espinosa, M. Almassalkhi, P. Hines, and J. Frolik, "System properties of packetized energy management for aggregated diverse resources," in *Proc. Power Syst. Comput. Conf.*, Jun. 2018, pp. 1–7.

[19] L. A. Duffaut Espinosa and M. Almassalkhi, "A packetized energy management macromodel with quality of service guarantees for demand-side resources," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 3660–3670, Sep. 2020.

[20] R. Malhame and C.-Y. Chong, "Electric load model synthesis by diffusion approximation of a high-order hybrid-state stochastic system," *IEEE Trans. Autom. Control*, vol. 30, no. 9, pp. 854–860, Sep. 1985.

[21] A. C. Kizilkale and R. P. Malhame, "Mean field based control of power system dispersed energy storage devices for peak load relief," in *Proc. 52nd IEEE Conf. Decis. Control*, Dec. 2013, pp. 4971–4976.

[22] W. Zhang, J. Lian, C.-Y. Chang, and K. Kalsi, "Aggregated modeling and control of air conditioning loads for demand response," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4655–4664, Nov. 2013.

[23] J. L. Mathieu, M. Kamgarpour, J. Lygeros, G. Andersson, and D. S. Callaway, "Arbitraging intraday wholesale energy market prices with aggregations of thermostatic loads," *IEEE Trans. Power Syst.*, vol. 30, no. 2, pp. 763–772, Mar. 2015.

[24] G. S. Ledva, E. Vrettos, S. Mastellone, G. Andersson, and J. L. Mathieu, "Managing communication delays and model error in demand response for frequency regulation," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1299–1308, Mar. 2018.

[25] Y. Chen, A. Busic, and S. P. Meyn, "State estimation for the individual and the population in mean field control with application to demand dispatch," *IEEE Trans. Autom. Control*, vol. 62, no. 3, pp. 1138–1149, Mar. 2017.

[26] Y. Chen, A. Busic, and S. Meyn, "Estimation and control of quality of service in demand dispatch," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5348–5356, Sep. 2018.

[27] Y. Chen, M. U. Hashmi, J. Mathias, A. Bušić, and S. Meyn, *Distributed Control Design for Balancing the Grid Using Flexible Loads*. New York, NY, USA: Springer, 2018, pp. 383–411.

[28] N. Cammardella, J. Mathias, M. Kiener, A. Busic, and S. Meyn, "Balancing California's grid without batteries," in *Proc. IEEE Conf. Decis. Control (CDC)*, Dec. 2018, pp. 7314–7321.

[29] S. H. Tindemans, V. Trovato, and G. Strbac, "Decentralized control of thermostatic loads for flexible demand response," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 5, pp. 1685–1700, Sep. 2015.

[30] B. Zhang and J. Baillieul, "Control and communication protocols based on packetized direct load control in smart building microgrids," *Proc. IEEE*, vol. 104, no. 4, pp. 837–857, Apr. 2016.

[31] S. P. Nandanoori, S. Kundu, D. Vrabie, K. Kalsi, and J. Lian, "Prioritized threshold allocation for distributed frequency response," in *Proc. IEEE Conf. Control Technol. Appl. (CCTA)*, Aug. 2018, pp. 237–244.

[32] S. Bashash and K. L. Lee, "Automatic coordination of Internet-connected thermostats for power balancing and frequency control in smart microgrids," *Energies*, vol. 12, no. 10, p. 1936, May 2019.

[33] S. Esmaeil Zadeh Soudjani and A. Abate, "Aggregation and control of populations of thermostatically controlled loads by formal abstractions," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 975–990, May 2015.

[34] R. Hermans, M. Almassalkhi, and I. Hiskens, "Incentive-based coordinated charging control of plug-in electric vehicles at the distribution-transformer level," in *Proc. Amer. Control Conf. (ACC)*, Jun. 2012, pp. 264–269.

[35] Resilient Home. *Green Mountain Power*. Accessed: Nov. 26, 2020. [Online]. Available: https://greenmountainpower.com/rebates-programs/home-energy-storage/resilient-home/

[36] M. Huang, P. E. Caines, and R. P. Malhame, "Individual and mass behaviour in large population stochastic wireless power control problems: Centralized and Nash equilibrium solutions," in *Proc. 42nd IEEE Int. Conf. Decis. Control*, Dec. 2003, pp. 98–103.

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

[38] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Aggregate flexibility of thermostatically controlled loads," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 189–198, Jan. 2015.

[39] J. Mathias, A. Busic, and S. Meyn, "Demand dispatch with heterogeneous intelligent loads," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, Jan. 2017, pp. 3138–3147.

[40] ISO New England. *Simulated Automatic Generator Control (AGC) Setpoint Data: Energy Neutral*. Accessed: Oct. 18, 2017. [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/grid/-/tree/simulated-agc

[41] A. Baird, G. Huang, C. Munns, and O. Weinstein. (2018). *Serverless Architectures With AWD Lambda*. Disponível em. [Online]. Available: https://d1.awsstatic.com/whitepapers/serverless-architectures-with-aws-lambda.pdf

[42] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions*, vol. 165. New York, NY, USA: Wiley, 1997.

L. A. Duffaut Espinosa (Senior Member, IEEE) received the B.S. degree in physics from the Universidad Nacional de Ingeniería, Lima, Perú, in 2003, the M.S. degree in mathematics with mention in stochastic processes from Pontificia Universidad Católica del Perú, Lima, in 2005, and the Ph.D. degree in electrical and computer engineering from Old Dominion University, Norfolk, VA, USA, in 2009.

He is currently an Assistant Professor with the Department of Electrical and Biomedical Engineering, University of Vermont, Burlington, VT, USA. His research interests span modeling, control and estimation of nonlinear systems, stochastic processes and algebraic combinatorics.

Adil Khurram (Student Member, IEEE) received the B.S. degree from the National University of Sciences and Technology, Islamabad, Pakistan, in 2013, and the M.S. degree from the American University of Sharjah, Sharjah, UAE, in 2016, in electrical engineering. He is currently pursuing the Ph.D. degree in electrical engineering with the Department of Electrical and Biomedical Engineering, University of Vermont, Burlington, VT, USA.

His research interests include statistical electrical load modeling and control of distributed energy resources in power systems.

Mads Almassalkhi (Senior Member, IEEE) received the B.S. degree in electrical engineering with a dual major in applied mathematics from the University of Cincinnati, Cincinnati, OH, USA, in 2008, and the M.S. degree in electrical engineering: systems and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA, in 2010 and 2013, respectively.

He is currently an Assistant Professor with the Department of Electrical and Biomedical Engineering, University of Vermont, Burlington, VT, USA. His research interests span multi-timescale control of DERs, energy optimization in power systems, and multienergy systems.