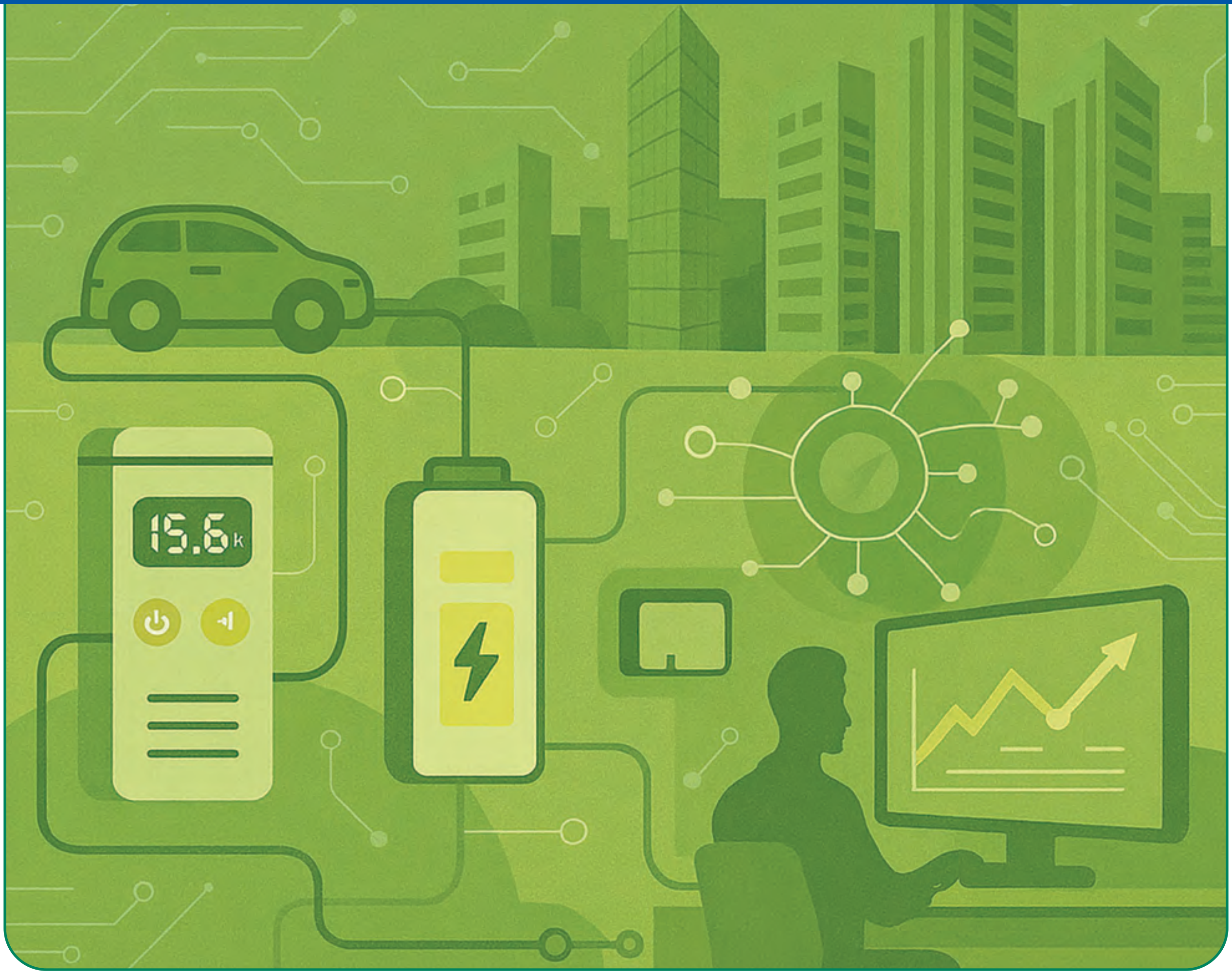
FEATURE

NAVIGATING THE PHYSICS OF VIRTUAL POWER PLANTS

The Role of Devices, Controller Design, and the Grid

Mads R. Almassalkhi, Paul Hines, Mazen Elsaadany, and Hamid R. Ossareh



(SOURCE: GRAPHICS CREATED USING AI-GENERATION. FOR IMAGE CREDITS, PLEASE SEE THE "ACKNOWLEDGMENT" SECTION OF THIS ARTICLE.)

VIRTUAL POWER PLANTS (VPPs) AGGREgate distributed energy resources (DERs), such as batteries, rooftop solar, electric vehicles (EVs), and flexible loads, to deliver valuable grid services across different timescales. These services include peak load reduction, frequency regulation, energy arbitrage, and voltage support, all with the goal of improving system resilience, integrating renewable generation, and enhancing the utilization of existing infrastructure cost-effectively. This article discusses the fundamental physics that underpin VPP operations, highlighting how device and grid physics, and cyberphysical system architecture decisions, shape a VPP's capabilities. Understanding

Digital Object Identifier 10.1109/MPE.2025.3587661
Date of current version: 20 October 2025

these physical foundations is key to designing VPPs at scale and allowing them to realize their full potential in supporting a reliable, flexible, and decarbonized grid.

Understanding these physical foundations is key to designing VPPs at scale and allowing them to realize their full potential in supporting a reliable, flexible, and decarbonized grid.

We begin by defining flexibility clearly in terms of power capacity, energy availability, and responsiveness, drawing parallels between traditional physical power plants and VPPs. Next, we identify three key physical constraints on VPP flexibility: 1) device-level limitations, such as power ratings and quality-of-service (QoS) constraints; 2) cyber-physical coordination limitations, including inherent challenges posed by communication latency and feedback signals; and 3) grid-level constraints, where physical transformer, cable, and voltage limits can restrict DER contributions.

To address these bottlenecks, we explore two general classes of QoS-aware VPP implementations that can preserve flexible loads' local objectives within predefined limits (e.g., comfort and convenience) while delivering aggregate flexibility for different grid services. We then zoom out and discuss how the VPP Operators may be integrated into future grid-aware architectures that could be utility-[1]or VPP-centric, thus enhancing VPP performance at scale without compromising grid reliability. We also share practical perspectives from the industry VPP operator EnergyHub, whose real-world experiences demonstrate both the challenges and opportunities associated with managing DERs at scale. Through interesting use-cases and state-of-the-art academic research, this article outlines technical, yet practical approaches for overcoming limitations, and provides actionable strategies to facilitate deeper integration of VPPs into modern power systems.

What Do We Mean by Flexibility?

For decades, the bulk power system reliability and resilience has depended on flexibility from physical power plants (e.g., thermal generators, like coal and natural gas power plants) to engineer a reliable and resilient power system. For example, primary and secondary frequency control loops have historically depended on thermal generators' ability to respond to frequency deviations (i.e., disturbances) by changing their power outputs via governor droop (primary, proportional controller) or automatic generation control (secondary, integral control). Thus, today, the flexibility available from a physical power plant is well understood to be characterized by the:

- ✔ power outputs that can be reached (magnitude; megawatts)
- ✔ duration over which it can hold this power output (duration; hours)
- ✔ ramp-rate capability (MW/s).

The flexibility from physical power plants makes up about 90% of all power system-wide flexibility today and has contributed to grid reliability for decades. The availability of physical plants to provide flexibility is less critical in this context but is largely driven by factors related to outages (e.g., Texas winter storm *Uri*) and weather conditions (e.g., very hot or very cold days can limit generation capacity).

Instead of a single physical power plant, consider the flexibility available from a single 1-MW/1-MWh lithium-ion battery: It can reach 1 MW and hold it for up to 1 h, depending on its state of charge (SoC, available energy). Or, it can reach and hold 12 MW for up to 5 min or reach and hold 0.1 MW for up to 10 h. The battery's inverter can ramp up/down incredibly fast, which means their flexibility can be delivered across (almost) all timescales, which is why batteries today are being deployed to deliver many grid services, including fast frequency response (subsecond) and regulation (seconds),

[1] We use the terms *distribution system operator* and *utility* interchangeably, recognizing that regulatory and market structures are helpful to define them precisely.

This article outlines technical, yet practical approaches for overcoming limitations and provides actionable strategies to facilitate deeper integration of VPPs into modern power systems.

energy arbitrary (minutes), and peak reduction (hours). Summarizing this battery discussion, Figure 1 characterizes the *reach-and-hold set* of a single 1-MW/1-MWh (1C) battery at 50% energy SoC and shows the power values that can be reached and for how long we can hold this power. Clearly, if the SoC were 0% (100%), the hold time for charging (discharging) would double while the discharging (charging) hold time would be effectively zero.

Now, what kind of flexibility can residential appliances offer? Thermostatically controlled loads (TCLs), such as electric water heaters or air conditioners, and deferrable loads like EV chargers, are not like batteries or conventional power plants. These devices typically consume large amounts of energy over time and also draw high power in short bursts (e.g., more than 2–3 kW during a 20-min heating or cooling cycle). Their energy use is also closely tied to complex human behaviors and QoS needs, such as expecting hot water during a shower, maintaining home comfort on a hot day, or ensuring a full EV battery before a long family road trip. In addition, devices already have local control logic that predefines how a device responds to its changing need for energy (NFE) to ensure QoS requirements are met: TCLs cycle on and off based on local temperature conditions being met. Thus, given the complexity of QoS and local control logic, the simplest way to "coordinate" such devices historically has been to voluntarily and temporarily switch off all devices during specific hours of the year as part of a residential demand–response (DR) programs, where peak demand reduction (curtailment) provides significant value to utilities 5–15 times per year. Of course, when these devices are switched off for hours, they do not heat water, cool the air, or charge the EV, which means that the utility's desired load reduction also reduces the QoS for end-consumers, which can lead to unsatisfied DR participants. And since enrollment is voluntary and financial incentives are small (e.g., US$20 to US$100/y in electric bill credits in California's Power Saver Program), unsatisfied consumers may opt out *en masse* each year, which raises concerns about the long-term viability and scalability of DR programs that do not explicitly account for the required QoS.

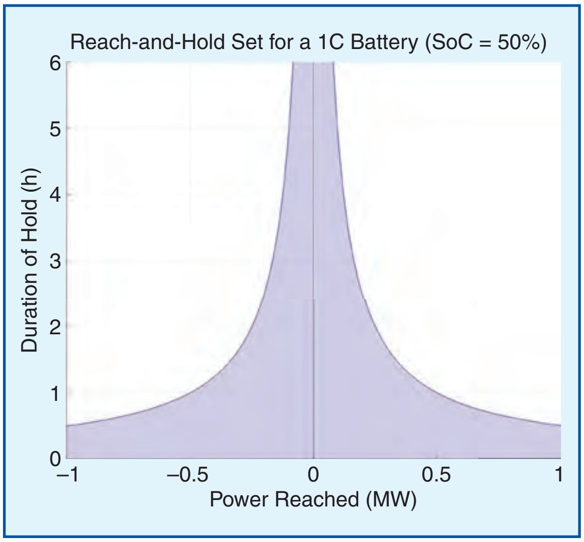


Figure 1. The reach-and-hold set for a 1 MW/1 MWh battery at 50% SoC showing which power outputs can be reached and for how long the power output can be sustained.

Furthermore, while DR programs are valuable today, they often underutilize the available flexibility by only making use of duration and magnitude aspects of flexibility and ignoring ramping capability. See for example Figure 2, where a fleet of 1,000 simulated air conditioners have their temperature setpoint increased by +2 °C as part of a DR event that reaches a 3-MW demand reduction and is held for about 2 h. By ignoring ramping, conventional DR programs today are poorly positioned to provide other increasingly valuable grid services on different timescales (temporal granularity), such as frequency regulation (seconds), price arbitrage (minutes), and operating envelopes (minutes). In addition, VPPs could respond to transmission grid operators and contribute to postcontingency and active risk mitigation schemes. For example, If a critical transmission line or generator fails, a VPP could be triggered to perform certain actions, such as discharging batteries or unloading the transmission system by reducing high-voltage ac load. Thus, there is significant value in unlocking flexibility from DERs. But this requires careful design and implementation of coordination schemes that optimize the aggregate, flexibility against local QoS requirements. Finally, at scale or in specific grid locations, flexibility will be needed to navigate binding physical grid constraints (or bottlenecks), which requires device coordination with spatial granularity (i.e., location matters).

The remainder of this article will, therefore, discuss how flexible devices can be coordinated to engineer a VPP that unlocks its full potential by accounting for the:

✔ limits of available flexibility

✔ QoS-aware coordination schemes across different timescales

✔ grid-aware coordination and the role of network limits.

We will also highlight some of these technologies with real case studies on dynamic load shaping and grid-aware coordination from one of the largest DER management system (DERMS) providers in the United States, EnergyHub.

The Role of Device Physics on VPP Flexibility

We have discussed the available flexibility from physical power plants and the reach-and-hold set of

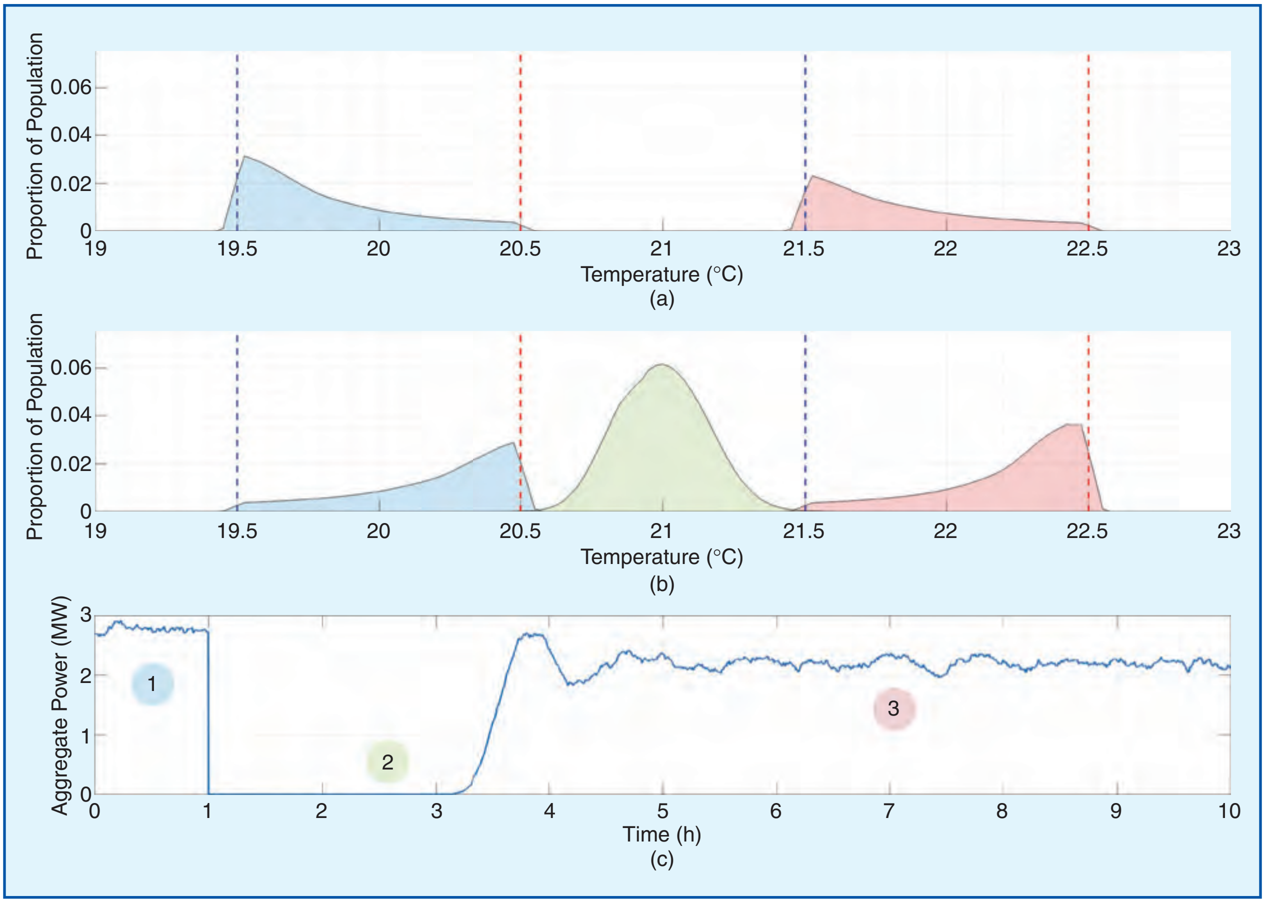


Figure 2. Three snapshots of the population dynamics of a fleet of simulated air conditioners with temperature setpoint of 20 °C, whose temperature setpoint is increased by +2 °C (a) and (b) and the effects on aggregate power over time (c). At ①, the ON and OFF temperature distributions are stationary (a in blue) with the aggregate power (relatively) fixed at 3 MW. A change in temperature setpoint is applied at hour 1, which causes all ON devices to switch OFF. The mean population temperature starts increasing at ② in green, dropping the total demand to zero. As the temperature of the OFF population keeps increasing, the OFF population eventually reaches the new temperature setpoint of 22 °C and achieves its new stationary temperature distribution at ③ in red. The DR event reaches and holds a 3 MW demand reduction for over two hours.

a battery (in Figure 1), but what can we say about the reach-and-hold set of a VPP composed of TCLs (e.g., air conditioners) on a hot day? To start to answer this question, consider the same VPP of 1,000 air conditioners from Figure 2 and assume that each TCL has a local exogenous disturbance acting on it that is well-modeled by a stochastic process (e.g., white noise).

Now, instead of increasing the temperature setpoint of all devices by +2 °C at the same time (e.g., start of DR event), what if we could dynamically schedule which proportion of devices adjust their setpoints—and when—to shape the aggregate power trajectory and achieve a desired reach-and-hold response from the VPP? That is, can we determine the full range of fixed power reductions (increases) and the corresponding durations for which the VPP can maintain such power change below (above) the level? The answer is: Yes, we can. We do this by building (linear, Markov-based) models of the original unchanged and changed setpoint population dynamics from the classic second-order equivalent thermal parameter model of a TCL. Then, we enforce a proportional setpoint control policy to couple the two dynamical population models, i.e., the proportion of the original population that change their setpoint are moved from the unchanged setpoint to the changed setpoint population. Then, we embed these (linear) models and controllers inside a multiperiod optimization problem that maximizes the reachable power reduction over a desired minimum hold duration, maximizing reach for a given minimum hold, which becomes the VPP's reach-and-hold set.

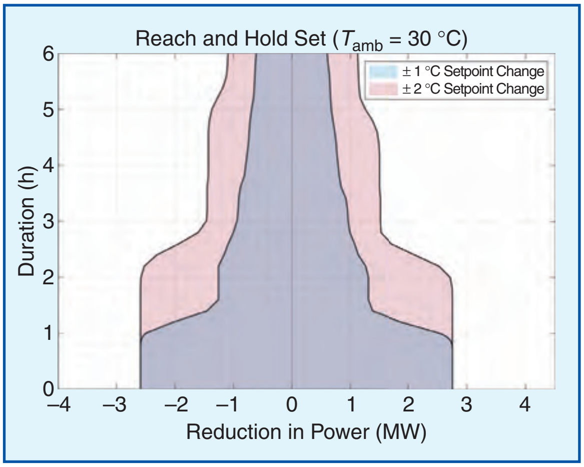


Figure 3. Characterizing flexibility of a VPP composed from 1,000 air conditioners via its reach-and-hold set, which is computed based on modeling and optimization with a ±1 °C and ±2 °C change in the air conditioners' setpoint.

Note that the reach-and-hold set described previously focuses on power reach values (megawatts) and power hold durations (hours) and does not directly consider ramp-rates or responsiveness (MW/s) explicitly. This is because we assume that air conditioners that are asked to change their setpoint do so much faster than the duration of the DR event. This translates into unbounded ramp up/down capabilities for the VPP similar to that of batteries. However, in practice, the ramp-up and -down capabilities of a TCL fleet is dictated by how fast the desired setpoint schedule can be computed at and communicated from the VPP (in the cloud) to the individual devices (in the field). Practically, communication and compute delays, and even actuation delays by TCL reacting to the new thermostat setpoint change, can limit the ramp up/down capability of VPP and is typically on the order of minutes. To overcome this minor practical issue, the VPP operator could start the DR event 5–10 min earlier, if necessary, to operate robustly.

We envision that these types of reach-and-hold sets can be developed into adaptive playbooks for VPPs, so utilities can easily optimize their different resources to achieve a desired aggregate DR response. For example, Figure 3 shows the corresponding reach and hold sets for a VPP composed of 1,000 air conditioners undergoing temperature setpoint changes of +2 °C (power reduction on right) and –2 °C (power increase on left). The ±2 °C change is compared to the reach and hold sets resulting from the same population undergoing +1 °C and –1 °C setpoint changes. While the ±1 °C changes exhibit similar power reach capabilities over an hour, its hold times are significantly less, which translates into a reduced energy capacity of the VPP compared to ±2 °C case. Interestingly, Figure 3 shows that for 30 °C ambient conditions and a +2 °C setpoint change, we can hold a 1.5-MW power reduction for 3 h (i.e., 4.5-MWh energy remaining in VPP before it runs out of energy), but for a –2 °C change, we can only hold a 1.5-MW load increase for 2.5 h (i.e., 3.8 MWh can be added before VPP is fully "charged"). We can consider the 1.5-MW increase for 2.5 h as a precooling event to improve the shape of the VPP's reach and hold set, increasing the VPP's DR potential beyond its 1.5 MW for 3 h.

Unlike the battery's reach-and-hold set in Figure 1, the VPP's reach-and-hold set lacks symmetry in both the reachable power and their corresponding hold times. This asymmetry is due to different modeling parameters embedded in the Markov model, including ambient temperature, initial setpoint temperature, and TCL population parameters. In addition, the second-order equivalent thermal parameter dynamics associated with building mass temperatures affects holds times for the ±2 °C and ±1 °C asymmetrically.

Nonetheless, Figure 3 informs a VPP operator of what is possible and what is not possible for their VPP at a given time. For example, with a +2 °C setpoint change, there is no physical way to reach and hold a demand reduction of 2 MW for three or more hours when ambient conditions are 30 °C. However, this is in part because we are limiting ourselves to a simple (open-loop) setpoint schedule and are not considering the available responsiveness of the fleet that can be unlocked with real-time DER coordination that also needs to account for devices' QoS requirements. Thus, DER coordination is discussed next.

The Role of DER Coordination Schemes, QoS, and Responsiveness

For a VPP to provide more services than just DR, it requires synergizing computing, communication infrastructure, data processing, and feedback signals, which together define a VPP as a *coordination scheme*. There are many coordination schemes in the literature: peer-to-peer, consensus-based, direct optimization-based DER scheduling, and indirect approaches, such as transactive energy. Thus, it is beyond the scope of this article to discuss every method. However, in addition to the open-loop DR schedule described via reach-and-hold sets, there are generally two broad classes for VPPs coordinating DERs. These are illustrated by Figure 4 and are:

- ✔ top-down or broadcast-based (synchronous) coordination
- ✔ bottom-up or device-driven (asynchronous) coordination.

Top-Down Coordination: A "Megaphone" Commanding All Devices, All the Time

Top-down coordination schemes rely on the VPP operator to simultaneously broadcast control signals to all DERs, which effectively allows the VPP to ramp up/down the entire OFF/ON DER populations. Unlike direct load control, top-down coordination can also leverage local sensing and computing to allow DER to selectively filter the VPP's common broadcast signals, enabling

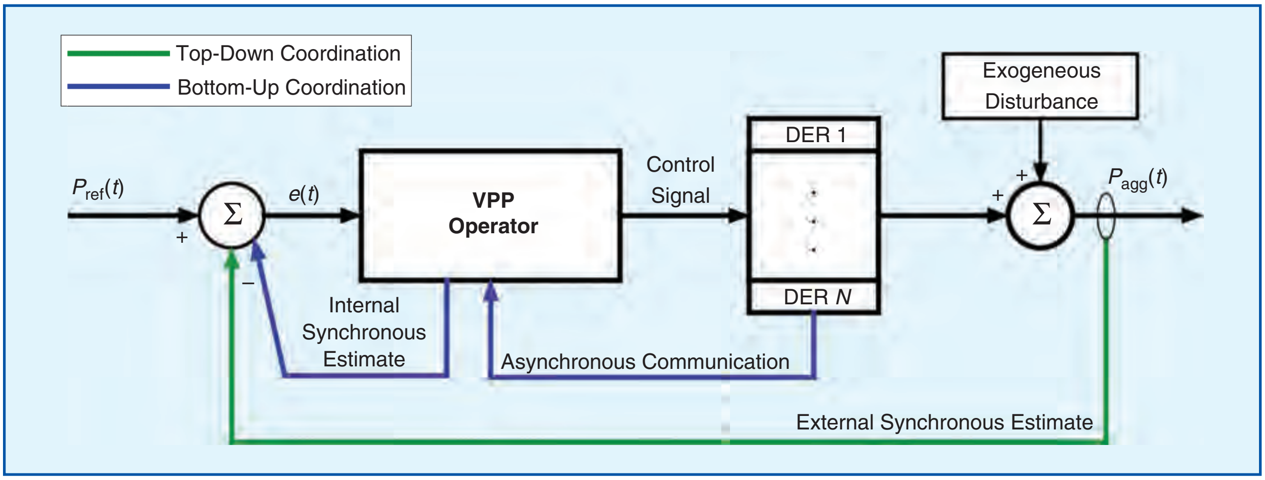


Figure 4. Top-down (green) and bottom-up (blue) represent two broad classes of VPPs.

differentiated device responses. For example, if a VPP ramp-up is required, the broadcast signal can be designed to prioritize turning on loads with higher NFE, such as air conditioners serving the warmer homes. However, differentiating DER responses based on live grid conditions is challenging for a top-down VPP, because it requires that: 1) the VPP operator has access to real-time grid information throughout the system and 2) the DERs can observe their local grid conditions. While the latter may be reasonable due to the integration of advanced sensing with inverter-based DERs, the former is not a good assumption today, unless the utility is the VPP operator and has visibility into distribution grid conditions. Thus, top-down VPPs (who are not the utility) will struggle to effectively manage grid congestion and keep the system operating within grid limits when DERs are at scale.

The inherent scalability of broadcast control and the availability of DER sensing and compute capabilities today means that top-down coordination schemes are well suited to unlock responsive VPPs. However, key limitations with top-down VPPs are their inability to 1) verify individual DER responses and 2) estimate the aggregate fleet demand in real time. That is, the green synchronous feedback signal in Figure 4 is generally hard to obtain in practice, which makes it difficult to implement feedback-based VPPs via top-down coordination. For example, if a VPP operator requires real-time power to be fed back from the fleet to generate its broadcast signal, all devices need to respond with measured power output all of the time, which imposes onerous communication requirements on implementation. Thus, top-down VPPs are often designed to rely on open-loop predictions or assume that total "system" demand can be inferred, e.g., from measured substation supervisory control and data acquisition or grid frequency. That assumption may be reasonable, if most demand in the system is enrolled within a single VPP.

Ultimately, while top-down VPPs offer scalability, unlock responsiveness, and can preserve QoS by leveraging local DER sensing and computing to enable selective response mechanisms, there are notable challenges due to the absence of accessible (synchronous) feedback signals and grid-awareness.

Bottom-Up Coordination: Always Listening to Some Devices, but Never All at the Same Time

In bottom-up (device-driven) coordination schemes, the VPP operator does not broadcast commands to all devices all of the time, but instead "listens" and responds to incoming, asynchronous, event-based DER communications (see blue signals in Figure 4) that implicitly can embed QoS-aware and grid-aware capabilities into the VPP operations.

For DERs equipped with local sensing and computing capabilities, the bottom-up VPP operator endows each DER with local control logic that defines if and how often a DER will asynchronously interact with the VPP, e.g., as a function of the DER's NFE. The DER owner then provides the necessary input parameters via the VPP operator's or distribution system operator's (DSO's) app to specify the NFE as a function of the DER's state variable (e.g., desired temperature setpoint or time until departure with the SoC). These DER-to-VPP communications can include requested actions, such as a desire to switch ON/OFF for a prespecified epoch (e.g., 5 min), along with relevant device parameters, like power capacity and binding DER operational constraints that help the VPP construct a real time, synchronous estimate of aggregate VPP power. The DERs could also compute and communicate a local "fitness" value to indicate near-term available flexibility, allowing the coordinator to prioritize the most "fit" DERs. Nonetheless, because bottom-up device interactions with the VPP operator are largely driven by QoS metrics (i.e., the higher the NFE, the more likely a DER will communicate their request to turn ON/OFF to the VPP), only a small subset of DERs will communicate with the VPP's any given window of time. These communicated ON/OFF requests also define the VPP operators capacity to ramp-up or down. Unlike the top-down VPP, the bottom-up VPPs will not be able to ramp-up/down with entire OFF/ON populations relying on round-trip communications between DERs making requests and the VPP communicating responses, which begets additional lag on the order of 0.5–10 s and reduces the ramp-rate limits a bit. Nonetheless, the asynchronous nature of bottom-up requests means that each DER operates on its own schedule, which promotes a low communication overhead and scalability. Furthermore, the DER requests can be common (e.g., charge/discharge) across different DER types, which simplifies VPP operator decision logic and enables real-time processing of these incoming device requests. Importantly, because the coordinator directly receives and responds to all DER communications, it can accurately estimate the aggregate VPP power output in real time, which enables closed-loop VPP control.

Finally, since DERs are communicating "upstream" in bottom-up coordination schemes, they can use local grid information (e.g., substation tag or measured local voltage magnitude) to inform their DER request logic or share the data with the VPP, which can unlock a grid-aware VPP operator. Such a bottom-up scheme would allow a VPP operator to automatically adapt DER coordination to live grid conditions without ever having direct access to live grid conditions. By leveraging live grid information and avoiding overloading distribution feeders (transformers, cables, voltages), a bottom-up VPP scheme can actively enhance grid reliability with DERs deployed at scale, thereby deferring or avoiding capital expenditures and supporting sustainable rate design.

The Role of Grid Physics, Data, and Constraints

If we endow VPPs with grid-aware capabilities, we can dynamically navigate DERs around grid bottlenecks to avoid overloading transformers and cables.

We have already discussed grid-aware coordination from the perspective of how top-down and bottom-up coordination schemes may integrate grid information. Of course, distribution grid congestion from VPP actions are mostly a concern when DERs are deployed at sufficient density in a specific region of the grid (e.g., affluent neighborhoods with more EVs). Thus, VPP coordination of DERs without awareness of local grid conditions poses significant risks to reliability in the (near) future. These reliability risks will either translate into blackouts (unlikely outcome) or costly grid upgrades (much more likely), impacting energy affordability. Conversely, if we endow VPPs with grid-aware capabilities, we can dynamically navigate DERs around grid bottlenecks to avoid overloading transformers and cables and mitigate feeder voltage violations. Thus, a grid-aware VPP operator would help utilities or DSOs to defer or avoid grid upgrades, which helps manage ratepayer costs long term. Of course, integrating (live) grid data with VPP operators is not straightforward, unless the DSO also acts as the VPP operator and has sufficient visibility into its feeders. Thus, grid-aware VPP operators give rise to interesting technical questions:

- ✔ Should utilities share their grid data with active VPP operators in their area?
- ✔ With grid information available, what if one VPP operator dispatches DERs in a way that effectively blocks other VPPs?
- ✔ With multiple VPP operators active in a feeder, how does a utility mitigate the risk of undesirable coincident dispatches that impact grid reliability?
- ✔ Given grid reliability and cybersecurity concerns, should utility be required to manage DERs?

Questions about data sharing between the DSO and VPP operators require careful consideration of the underlying grid architectures, which are summarized in Figure 5 and include today's common *Total transmission system operator (TSO)* architecture [Figure 5(a)], where the VPP operators are unaware of distribution grid conditions, focus on wholesale market opportunities, and only interact with utility on DR services.

As an alternative to the Total TSO, one can adopt a more utility-centric Total DSO architecture, as shown in Figure 5(b), where it becomes the utility's role to site, deploy, and dispatch DERs using either their own internal DER coordinator or integrate with a DERMS platform to unlock grid-aware DER coordination. The Total DSO's deep integration of DERs is less suitable in a deregulated context since DERs become part of distribution grid planning and operations, which impacts integrated resource planning outcomes long term and treats the DERs as part of a utility's poles-and-wire infrastructure (e.g., as distributed capacity). Overall, the Total DSO approach aligns DERs with grid reliability objectives, enabling additional system value, particularly when paired with local services, such as volt-var optimization from smart inverters. However, realizing these benefits requires a regulatory framework that supports utility-led DER integration and values distribution-level services.

In Figure 5(c), we have a Hybrid DSO architecture that enables utilities and VPP operators to cooperate on reliability and grid-aware DER coordination while enabling DER owners to also benefit from market-facing objectives. The cooperation is achieved by leveraging common communication protocols, such as IEEE 2030.5 and 2030.11 for DERMS implementations. One promising hybrid approach to enable grid-aware VPP operations is the DSO's implementation of dynamic operating envelopes (DOEs), within which VPPs are allowed to operate freely as long as they do not exceed certain (nodal) power limits. Of course, these (nodal) power limits reflect the physical limits of the grid and need to be computed from an accurate grid model (or data-driven process), which is still an active area of research. As with any limits, DOEs will constrain (a portion of) the VPP's available flexibility at various times. DOEs require secure data exchange between DSOs and VPP operators to minimize cybersecurity risks through encrypted communication, strict access controls, and adherence to standards, such as SOC 2 and ISO 27001. These protections enable the implementation of software- defined dispatch limits that are trusted by both the DSO and participating VPPs.

To ensure trust and accountability within these architectures, modern VPP programs rely on rigorous evaluation, measurement, and verification processes using advanced metering infrastructure and DER telemetry to assess DER performance. Programs like Massachusetts' ConnectedSolutions

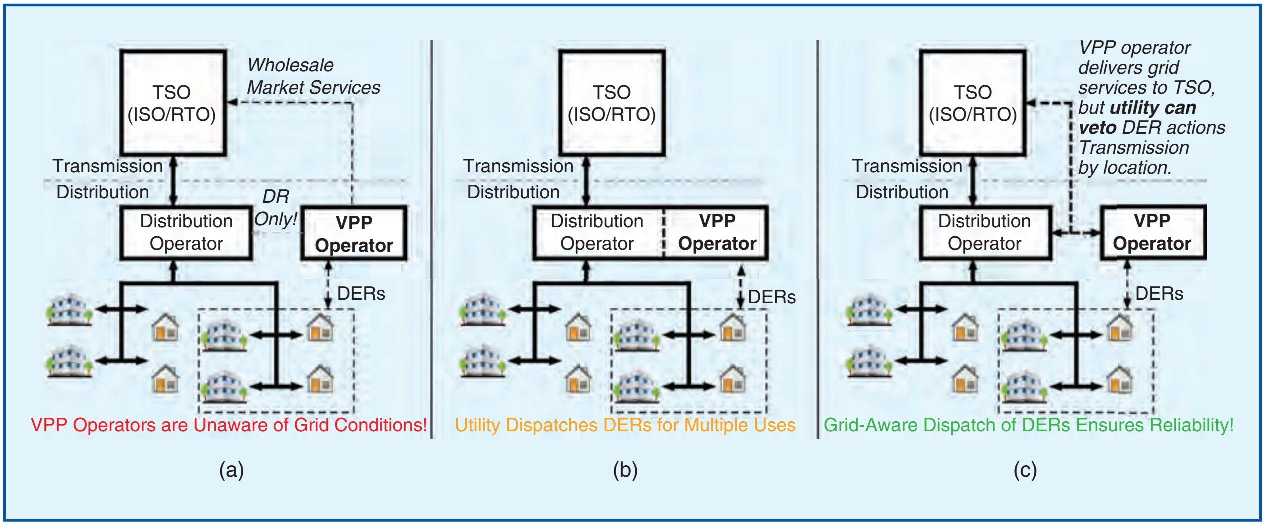


Figure 5. A summary of relevant grid-DER control architectures that consider different ways that VPP operators can interact with the DSO. (a) The Total TSO has VPPs separate from DSO, except for DR, because DERs are not yet at scale in many places, so have no regional grid impacts. (b) At high DER penetrations, the Total DSO benefits from dispatching DERs for the sake of enhancing reliability within the DSO's system rather than DERs mainly pursuing wholesale market objectives. (c) Future interactions are likely to take on a hybrid approach, where VPP operators can dispatch DERs for wholesale market services, but subject to a DSO veto that ensures a grid-aware, reliable dispatch.

demonstrate how well-designed incentives—such as fixed payments for smart thermostats and pay-for-performance for batteries—can align DER behavior with grid needs without requiring DSO ownership. These verifiable, data-driven structures strengthen participation, regulatory confidence, and program effectiveness. Next, we bring together the discussion so far into three real-world case studies.

Case Study 1: Dynamic Load Shaping With Thousands of Thermostats

Many utilities would like to have more precise control over load shapes, relative to what one would get from a conventional all-OFF DR event, which often generates large amounts of load reduction at the beginning of an event period, and then less load reduction at the end of an event. A utility may, for example, want to control the load shape to counteract the "duck curve" that comes with growing adoption of rooftop solar photovoltaic. With that in mind, EnergyHub developed and tested a top-down DER coordination technology that allows a utility to choose the load shed shape that they prefer, as shown in Figure 6. Given the target VPP output (black line in Figure 6) and a fixed group of thermostats (colored lines in Figure 6), the technology then optimally dispatches the different groups by sending each group an open-loop schedule of temperature setpoint changes to deliver the aggregate target load shape.

Figure 6 represents real-world results from one test of this technology with about 10,000 thermostats. In this case, the target VPP output is negative from 1200 hours to 1400 hours to represent a pre-cooling strategy (with load increasing), but then during the event period from 1600 hours to 1900 hours, the different thermostat groups were optimally scheduled, so that there was less load shed in the first hour, more in the second hour, and then less in the third hour. The results clearly demonstrate that it is possible to shape the physical response of TCLs to achieve the desired reach-and-hold target outputs, even with a large aggregation of thermostats.

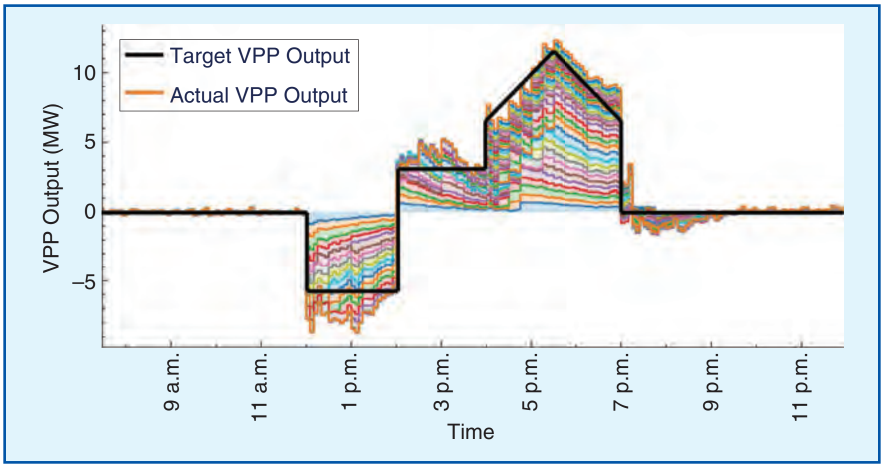


Figure 6. Dynamic load shaping test using a real-world VPP of 10,000 thermostatically controlled devices, organized into different groups (stacked colored lines) and strategically scheduled to track a target aggregate load shape (black line). Devices were scheduled to precool homes during midday solar surplus and then ramp to align with the evening peak. The net VPP output (outer orange line) closely follows the desired trajectory, with group-wise stacking revealing how device physics shapes aggregate flexibility.

Case Study 2: Bottom-Up VPP Delivers Frequency Regulation With 200+ Loads

A new VPP technology—implemented in bottom-up fashion—was commercialized by grid-tech startup, Packetized Energy, which launched in 2016 and was acquired by EnergyHub recently. The bottom-up VPP technology, Packetized Energy Management (PEM), endows DERs with enough local intelligence to track their NFE and then enables each DER to request an energy packet asynchronously to the VPP operator, which is either accepted (allowing the DER to consume power for a fixed epoch of 1–10 min) or rejected (causing the DER to not consume power). The request rate is designed to be proportional with the NFE, which means that the higher the DER's NFE, the less time (on average) between the DER's requests, which indirectly prioritizes devices with high NFE (over time) and helps ensure QoS objectives across the DER fleet. The objective of the VPP operator is then simply to either accept or reject incoming

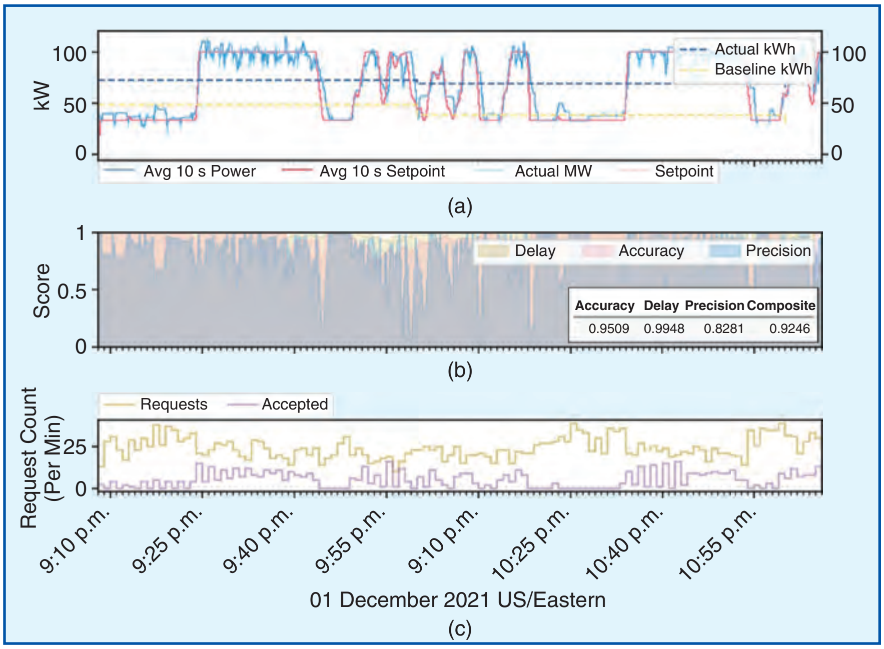


Figure 7. Packetized energy bottom-up VPP in the field coordinating 223 heterogeneous electric water heaters. (a) The VPP delivers frequency regulation over 120 min. (b) The composite score of 92% was representative across four different test days. (c) DER packet request rates and VPP operator's acceptance rates per minute.

requests to minimize the VPP's tracking error, *e(t)* from Figure 4. Clearly, if there are only a few requests within, say, a 10-s window, the VPP operator cannot ramp the fleet's demand up much over that time (since previously accepted packets are completed, which turns loads OFF and reduces aggregate demand). Thus, the request rates inform the VPP operator of available ramping capacity available to minimize power reference tracking errors. In fact, the effective bandwidth of the closed-loop PEM scheme is related to the packet length: The shorter the packets, the more packets are available to send a request, the better the VPP can deliver faster grid services. However, with shorter packets and more requests, the devices cycle ON/OFF more (on average) and communication overhead costs increase. Such tradeoffs are common for bottom-up implementations of VPPs.

The aforementioned bottom-up PEM technology was demonstrated for frequency regulation in December 2021 with a population of 223 flexible, heterogenous electric hot-water heater loads in Vermont and the Carolinas. Specifically, the frequency regulation test used PJM's Reg-D data, which is a form of fast frequency (power) regulation service that sends the VPP operator an updated normalized reference signal every 2–4 s, which is scaled by the regulation capacity offered by the VPP. Interestingly, PJM uses a pay-for-performance scheme to incentivize participation. The metric used to quantify performance is the composite score, which is the average of normalized precision, accuracy, and delay scores over the settled hour,[2] and must be greater than 75% for the resource to earn revenue during that hour. The VPP's demonstrated performance under PEM is shown in Figure 7, including aggregate VPP and reference setpoint power [Figure 7(a)], performance scores [Figure 7(b)], and asynchronous request rates shown per minute [Figure 7(c)]. The VPP scored a composite score of over 92% and a precision score of 83%. Most thermal generators score around 80%–90%, while utility-scale battery systems score 95% or better. This was possible with about 25 requests per minute, which is only about 0.2% of the population making a request per second on average. In 2021, this demonstration represented one of the first real-world implementations of a bottom-up VPP that could provide fast frequency regulation and other grid services.

Case Study 3: Grid-Aware Load Shaping With EVs

EVs have the potential to reduce emissions and costs of transportation, but the charging of EVs can add stress to aging grid infrastructure. Many have written about the potential to use various

[2] *Precision* measures the root mean square tracking error, *accuracy* reflects correlation with the reference, and *delay* captures the time shift yielding the highest correlation.

algorithms to simultaneously optimize for both grid economics and customer quality while also operating within power grid limits, but there have been few real-world demonstrations of this technology. This case study provides data in Figure 8 from an EnergyHub demonstration of this grid-aware technology for a VPP with 60 EVs as part of a managed charging program.

In this managed charging system, customers are operating under a simulated retail rate in which the peak-hour charging is from 0900 hours to 2100 hours. They use a mobile application to set the time at which they generally need their vehicle charged in the morning (e.g., 0700 hours), and a desired EV target charge level (e.g., 90%), which defines the EV's NFE. The utility operator will then organize the vehicles into multiple, potentially overlapping logical groups based on distribution network constraints: feeders, transformers, etc. Every 15 min, the EnergyHub DERMS software collects data from the vehicles and develops a charging plan for the fleet of vehicles to ensure that all EVs meet their NFE requirement (QoS), that vehicles are charged to minimize retail or wholesale cost of electric energy, and to keep the aggregated charging for each group of vehicles under the utility-specified group limits. The first step of this plan is then delivered to the vehicles in the form of stop/start charge commands. And then the process repeats.

Figure 8 shows results from the application of this method to a group of EVs that are managed top-down to charge during off-peak hours. The system can ensure that all customers obtain the charging that they need before they leave in the morning, while also keeping the aggregate load close to or below a (soft) grid-aware power limit. Some charging is allowed above the limit, if needed to satisfy customer quality of service requirements, which are treated as the priority for the charging algorithm.

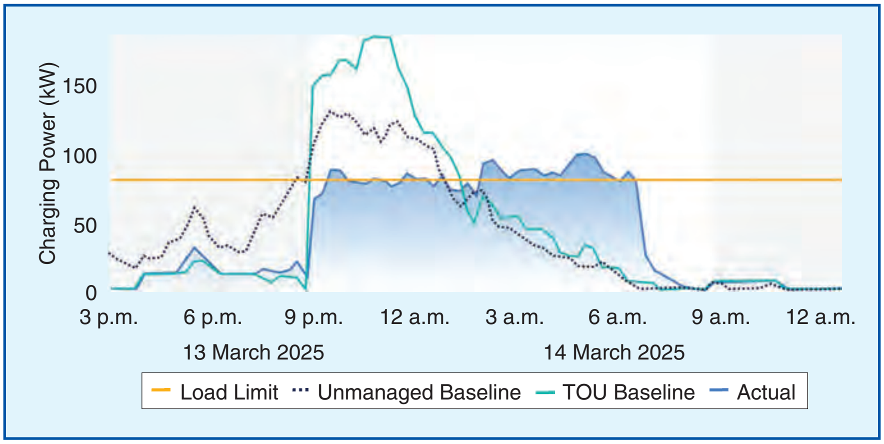


Figure 8. Grid-aware DER coordination. VPP manages the charging of a group of about 60 real EV. The objective is to shift charging away from peak periods (gray shade), while keeping aggregate load (blue shade) under the specified load limit (orange line). The turquoise line shows an estimate of the charging that would have occurred if the system had optimized purely for the time of use (TOU) rate while the black dashed (···) line shows the unmanaged baseline.

Conclusion

This article has examined the physical foundations that shape the flexibility VPPs can offer when coordinating DERs to provide grid services. We defined flexibility in terms of the power magnitude that VPPs can sustainably reach and hold for specific durations (e.g., DR events), as well as the (ramp) rates at which they can adjust their power output, which we termed *responsiveness*. Using this definition, we introduced the concept of reach-and-hold flexibility sets for VPPs, highlighting the critical importance of quantifying end-user QoS impacts when providing grid services.

We then discussed how DER coordination approaches—specifically top-down (broadcast-based) and bottom-up (device-driven)—influence the operational capabilities and responsiveness of VPPs. Moving beyond coordination methods, we further emphasized the significance of grid physics and the necessity of integrating real-time grid conditions into VPP operations to ensure grid reliability as DER deployments scale. Three real-world VPP case studies demonstrated these principles practically: QoS-aware coordination of 10,000 residential thermostats during a DR event, bottom-up frequency regulation using over 200 electric water heaters, and grid-aware management of a fleet of 60 EVs.

Looking ahead, several open challenges remain to fully unlock the potential of VPP flexibility and ensure effective QoS- and grid-aware VPP operations. One key challenge is facilitating improved data sharing and systematic operational collaboration between utilities and VPP operators. Another critical need is establishing standardized DER capabilities and communication protocols to streamline their integration into VPPs, maximizing device availability and operational efficiency.

Moreover, effectively managing the inherent latency, controllability, and observability limitations, and modeling uncertainties associated with DER coordination schemes remains essential. Addressing these challenges calls for innovations in optimization techniques, sensor and monitoring technologies, and dynamic, grid-aware coordination schemes and demonstrations. Such advancements will also enable accurate dynamic characterizations of VPPs' available flexibility, including ramp-rates, that reflect the underlying DER physics and coordination, enhancing VPP reliability and resilience, ultimately delivering scalable and robust grid services. Meeting these challenges will significantly benefit DER owners, grid operators, and broader grid resilience and decarbonization efforts alike.

Acknowledgment

The opener image was created using OpenAI's ChatGPT 4.0.

For Further Reading

J. L. Mathieu et al., "A new definition and research agenda for demand response in the distributed energy resource era," *IEEE Trans. Energy Markets, Policy Regulation*, early access, Mar. 26, 2025, doi: 10.1109/TEMPR.2025.3554734.

U.S. Department of Energy, "Pathways to commercial liftoff: Virtual power plants," Jan. 2025. [Online]. Available: https://www.constructive.org/liftoff

"Decoding DERMS: Options for the future of DER management," Smart Electric Power Alliance, Washington, DC, USA, Mar. 2025. [Online]. Available: https://sepapower.org/resource/decoding-derms-options-for-the-future-of-der-management/

P. De Martini, S. Succar, and P. Cook, "Sourcing distributed energy resources for distribution grid services," U.S. Department of Energy, Washington, DC, USA, Dec. 2024. [Online]. Available: https://www.energy.gov/sites/default/files/2024-12/Sourcing%20DER%20for%20Dist%20Services%20final%2012.17.24.pdf

"From obstacle to opportunity - How managed charging can mitigate the distribution impacts of EV charging." EnergyHub. Accessed: Jul. 17, 2025. [Online]. Available: https://www.energyhub.com/resource/from-obstacle-to-opportunity-how-managed-charging-can-mitigate-the-distribution-impacts-of-ev-charging

M. Almassalkhi et al., "Hierarchical, grid-aware, and economically optimal coordination of distributed energy resources in realistic distribution systems," *Energies*, vol. 13, no. 23, 2020, Art. no. 6399, doi: 10.3390/en13236399.

M. Almassalkhi, J. Frolik, and P. Hines, "Packetizing the power grid: The rules of the internet can also balance electricity supply and demand," *IEEE Spectr.*, vol. 59, no. 2, pp. 42–47, Feb. 2022, doi: 10.1109/MSPEC.2022.9706403.

Biographies

Mads R. Almassalkhi is with the Center for Resilient Energy and Autonomous Technologies in Engineering, University of Vermont, Burlington, VT 05403 USA.

Paul Hines is with EnergyHub, Brooklyn, NY 11217 USA.

Mazen Elsaadany is with the Center for Resilient Energy and Autonomous Technologies in Engineering, University of Vermont, Burlington, VT 05403 USA.

Hamid R. Ossareh is with the Center for Resilient Energy and Autonomous Technologies in Engineering, University of Vermont, Burlington, VT 05403 USA.