The Challenge of Integrating Demand Response in Capacity Remuneration Mechanisms Providing a Comprehensive Theoretical Framework

Pablo Rodilla, Paolo Mastropietro, Paulo Brito-Pereira

This article is a preprint. Please cite the published version:

10.1109/MPE.2023.3269551

THE CHALLENGE OF INTEGRATING DEMAND RESPONSE IN CAPACITY REMUNERATION MECHANISMS

Pablo Rodilla, Paolo Mastropietro, and Paulo Brito-Pereira

Capacity Remuneration Mechanisms (CRMs) have become a pillar of the design of decarbonising electricity markets. By complementing the economic signals conveyed by the energy market, they aim at enhancing resource adequacy, particularly in the current context in which power systems transition towards low-carbon technologies. They are also being mentioned as a key piece to prevent, in the future, scenarios such as the energy crisis that started in 2021 in the European Union.

Although CRMs have been frequently criticised and identified as a tool for subsidising conventional generation driven by fossil fuels, they have shown their potential in fostering new technologies and business models. International experiences have shown how demand response can compete with generation technologies and play a relevant role in capacity mechanisms. For instance, demand resources covered 10% of the capacity market in PJM (Pennsylvania, New Jersey, and Maryland power system, one of the largest interconnections in the United States) in recent years. Figure 1 shows how dependent demand-side response has been upon the revenues coming from the capacity market of this power system. Demand-response participation in European CRMs is also growing, but it accounts only for 3% of the demand for firm capacity in the region.

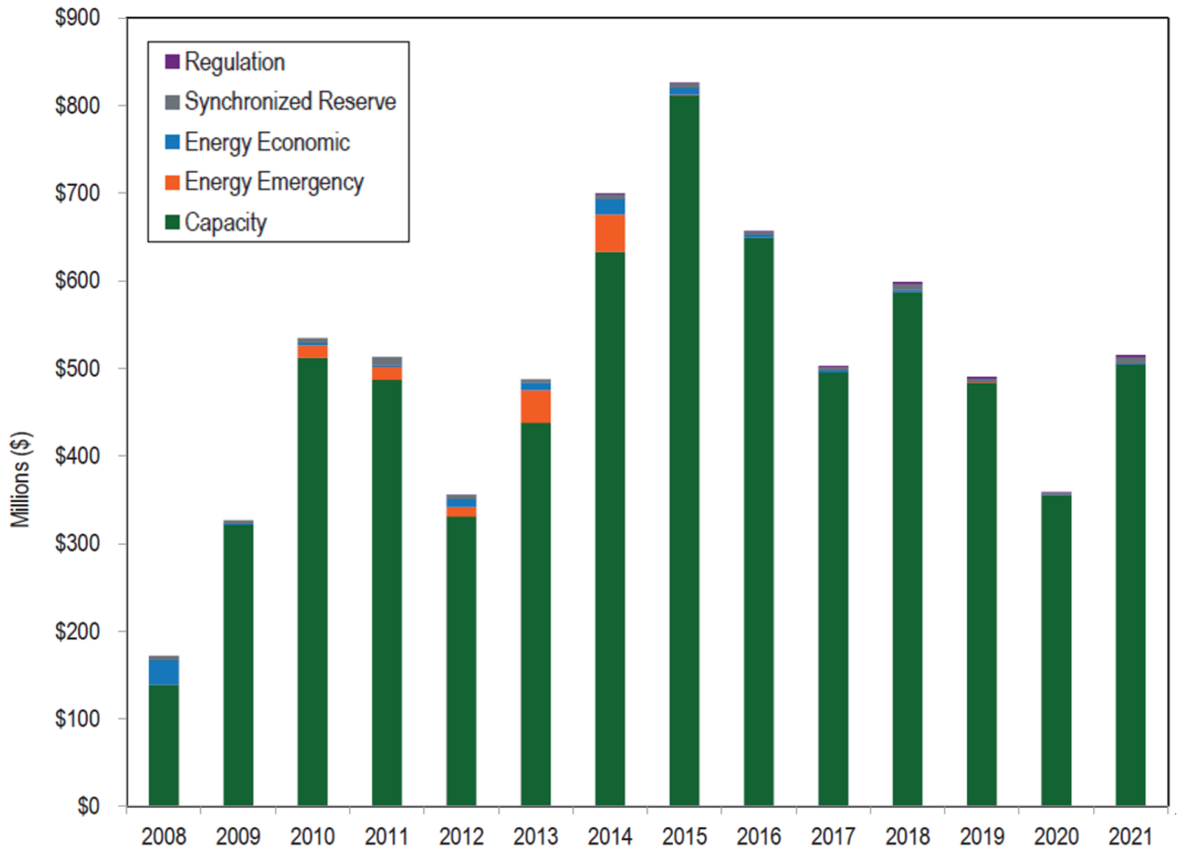

Figure 1. Evolution of demand resources income in PJM, chart from Monitoring Analytics' State of the Market Report

Integrating demand resources in CRMs is beneficial for the system, since it reduces overall costs and promotes resources whose contribution in terms of flexibility will be much needed in future power systems. However, this participation adds a layer of complexity to the design of capacity mechanisms. Two key elements in the design of the CRM are particularly relevant when it comes to integrating demand response: i) the way the demand to be covered by the capacity mechanism is defined by the regulator and ii) the methodology to allocate the costs of the CRM among consumers who benefit from that coverage. There is currently a gap in the academic literature on resource adequacy and CRMs, which has often missed delving into these two aspects.

The goal of this article is to define a comprehensive framework for the participation of demand resources in capacity mechanisms, identifying all potential participation modes and highlighting the inefficiencies that could arise from certain designs. The article first assesses the aforementioned two central design elements of CRMs for an efficient participation of demand resources, i.e., the methodologies to set the demand for firm supply and to allocate the costs of the CRM. Then it defines a classification of all potential participation modes, listing the benefits and the potential inefficiencies of each of them, but also highlighting which are being used in real CRMs. Finally, the article draws regulatory recommendations.

1 DEMAND FOR FIRM SUPPLY AND COST-ALLOCATION STRATEGY

Most of the literature on demand participation in capacity mechanisms focuses on how the demand can sell some sort of demand-response (DR) service in the CRM as a capacity provider and some associated features of this participation (baselining and de-rating of DR agents). However, the potential role of demand response conditions the process earlier on, since it should be considered from the very start, i.e., when the regulator or the system operator estimates the expected need for firm supply during the delivery period.

Firm supply, a concept that encompasses both firm capacity and firm energy, is the expected contribution of a resource during scarcity conditions in the system. In capacity-constrained systems, such as power systems on the East Coast of the United States or in Europe, firm supply is usually computed through a de-rating factor (or capacity credit) to be applied to the installed capacity of the resource. In most capacity markets, the demand for firm supply is set administratively, without any active role from consumers. In centralised capacity markets, this exercise results in a demand curve that is used in an auction, which tries to reflect some degree of elasticity. This elasicity is defined administratively too. In decentralised capacity markets, demand for firm supply is computed for each load-serving entity through an administrative methodology.

The most efficient way for demand to participate would be the involvement of end-users in this initial phase, letting consumers define their own demand for firm supply without administrative interventions; this demand for firm supply would become the upper limit of their consumption during scarcity conditions in the system. Alternatively, the regulator could estimate an initial requirement for each consumer or consumer group and then allow them to increase or decrease such value. These approaches would also simplify and improve the efficiency of the cost allocation, since such self-declared demand for firm supply is the best cost driver on which to apply CRM charges. This process would be symmetric to the de-rating process for generation resources. Each consumer (or consumer group/category) would pay the costs of the capacity mechanism according to the consumer's expected "negative" contribution to the reliability of the system (since procuring 1 MW of firm capacity entails contributing "negatively" for that amount).

Nevertheless, in most CRMs, the demand for firm supply is estimated in a very aggregated way. For instance, the demand curve in capacity auctions is defined through an estimation of the whole-system demand and its evolution in the future, potentially applying a least-worst regret approach (e.g., in the United Kingdom). This approach significantly constrains the kind of participation that can be expected by demand resources, as analysed in the next section.

As already mentioned, another design element of CRMs that affects demand participation is the methodology for the recovery of the costs of the mechanism. The latter arises from the signature of capacity contracts with reliability providers, selected either through a centralised auction or through bilateral contracts. It must be noted that, once these contracts are signed, their cost is a sunk cost. Therefore, a reduction of demand during scarcity conditions with respect to the estimation made to set the demand for firm supply does not reduce these costs, which must be paid even if the service is not activated. This condition equates CRM costs to other residual costs in the power sector, as residual network costs or policy costs. According to economic theory, residual costs should be allocated minimising distortions to the economically efficient signals defined to recover other costs.

In the case of capacity mechanisms, this least distortion could be achieved, in principle, through the application of a fixed charge per customer. This charge should reflect the contribution of each consumer to the demand for firm supply. If this information cannot be extracted from the process of setting this demand, then the charge could be proportional to the historical consumption of each consumer during scarcity conditions. A moving average over a certain number of years or delivery periods could be applied to historical consumptions. Other cost-allocation methodologies are possible, as the application of capacity charges on withdrawals made during actual scarcity conditions. As already mentioned, these withdrawals are not the real cost driver (reducing demand in the real time would not reduce the CRM costs, once they have been incurred). However, such a cost-allocation methodology may foster an efficient behaviour from consumers that could reduce future needs for further firm capacity, although it would also result in the under-recovery of CRM costs.

International practices, however, favour simple volumetric charges applied over a very large number of hours. For instance, Ireland applies a CRM charge to electricity suppliers according to the demand they serve in day hours (from 7:00 to 23:00) during the entire year. Italy recovers 70% of CRM costs through charges applied to energy withdrawals during the

500 “peak hours,” defined as those hours in which the system is more likely to suffer a stress event; the remaining part of CRM costs is recovered through a much lower charge applied during the rest of the hours of the year. These cost-allocation strategies are equivalent to socialising the sunk CRM costs, without providing any efficient signal to consumers, but without guaranteeing cost recovery either. This approach affects the different modes of demand participation, as analysed in the next section.

2 DIFFERENT MODES OF DEMAND PARTICIPATION IN CRMS

After defining the necessary background on the methodologies for estimating the demand for firm supply and the strategies for CRM cost allocation, it is possible to classify the participation for demand resources in capacity markets into different participation modes. Using the standard terminology on demand response, two broad categories are identified as follows:

- Explicit participation: Consumers explicitly take part in some phase of the capacity market and assume binding commitments. They can do that i) in the demand side of the capacity market, by defining their demand for firm supply or ii) in the supply side, selling demand-response services that are equated to the reliability services offered by generators. This participation mode is addressed in section 2.1.
- Implicit participation: Consumers do not explicitly participate in the capacity market and they do not assume any binding commitment to reduce their load. However, they react to CRM charges during its operation, modifying their demand to reduce their contribution to the coverage of CRM costs (and, if charges are designed properly, their contribution to scarcity conditions). This participation mode is addressed in section 2.2.

2.1 Explicit participation

2.1.1 Demand side (opt-in or opt-out)

Although very infrequently used in practice, the most obvious way to involve consumers in capacity mechanisms would be to conceive an active role for them in the calculation of the demand for firm supply (according to some authors, PJM may be moving in this direction). Ideally, consumers could be asked to define beforehand the capacity they expect to need and withdraw from the power system during future scarcity conditions. The selection of this demand for firm supply could be informed by some brief report from the system operator, with estimations on the number of stress events expected in the system and on the range of the charges, to which this capacity demand would be subject. This value would limit actual withdrawals in real time during stress events, since consumers would commit not to exceed that capacity demand. In power systems where smart meters have already been deployed, this approach could encompass the entire demand, including residential or regulated demand. In a few countries (e.g., Spain), consumers are already asked to specify different contracted capacities, e.g., for peak or valley hours, which are subject to different charges. Widening this approach to include resource adequacy would only require asking consumers to specify an additional contracted capacity that would be used to limit consumption during scarcity conditions (or to impose sanctions on the withdrawals exceeding it).

This theoretical approach would move the responsibility of defining the demand for firm supply fully on consumers' shoulders. Although technically feasible, this shift may be challenging from a regulatory and political point of view. However, there are other approaches that mimic this first alternative and partially achieve its benefits. For instance, the regulator or the system operator could estimate an aggregated demand for firm supply but compute a disaggregated estimation for certain consumer categories (e.g., large commercial or industrial end-users). The latter would then be given the chance of opting out, i.e., of reducing or directly setting to zero the demand for firm supply assigned to them. The opt-out would generate a commitment that allows the system operator to limit withdrawals during scarcity conditions, but it would also exempt the consumer from paying CRM charges for the opted-out capacity. This approach is represented graphically in Figure 2, for a centralised capacity auction, whose demand and supply curves and the corresponding market clearing are depicted in a price-quantity chart. The same reasoning could be applied, however, to decentralized capacity markets, in which the obligation for each load-serving entity could be reduced through an opt-out of some of its end-users.

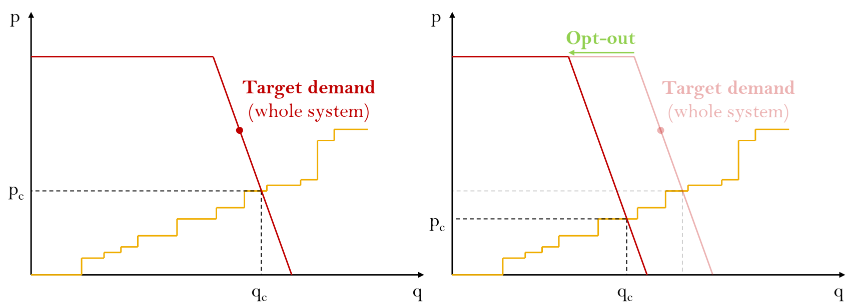

Figure 2. Explicit participation of demand resources in the demand side of the CRM through an opt-out (the demand curve mimics curves used in real CRMs, which represent demand elasticity)

A similar approach would consist in estimating the demand for firm supply only for certain consumer categories (e.g., residential or regulated demand). The rest of consumers would be required to define their own demand, through an explicit opt-in in the capacity market, which would generate the same commitments that have already been mentioned above. This opted-in capacity could be used to simply shift the demand curve (chart in the middle in Figure 3) or these consumers could be asked to present price-quantity demand bids, specifying also the value that they assign to the firm supply (chart at the right in Figure 3).

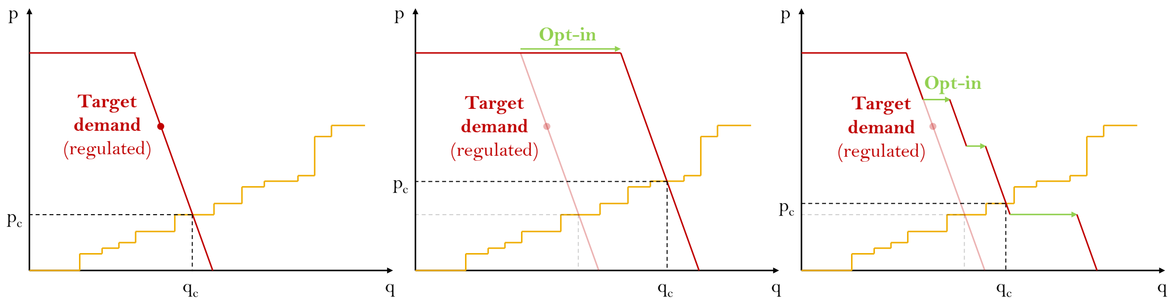

Figure 3. Explicit participation of demand resources in the demand side of the CRM through an opt-in

It must be noted that these approaches would also simplify the allocation of CRM costs. The demand for firm supply is the real driver of these costs. Therefore, if some or all consumers have a certain demand for firm supply earmarked to them, either estimated by the regulator/system operator or self-defined by the end-user, CRM charges could be easily applied to this capacity during each delivery period.

2.1.2 Supply side (demand response)

The disaggregation of the demand for firm supply is a complex task that, as mentioned in section 2, is hardly found in real CRMs. Especially in centralised capacity markets, the demand is commonly defined for the entire system and no opt-in or opt-out is allowed. In this context, consumers can still participate in the capacity market by offering demand-response services. These services are offered through price-quantity supply bids that go into the supply curve of the market, as shown in Figure 4. However, it must be highlighted that the consumers involved in such demand response are represented twice in the auction, both in the demand curve (since they are part of the whole-system demand for firm supply) and in the supply curve. This feature is prone to arbitrages and other inefficiencies, as analysed next.

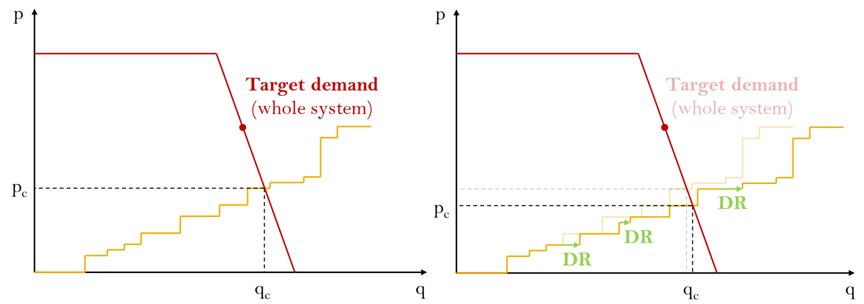

Figure 4. Explicit participation of demand resources in the supply side of the CRM through DR services

For consumers to offer demand-response services in the CRM, the regulator must design a reliability product that these resources will be allowed to trade in the capacity market. In principle, the reliability product should be the same for all the resources competing in this market segment and should reflect the ability of each agent to contribute to the reliability target in force in the power system. However, many regulators, both in Europe and the United States, have defined specific reliability products that are tailored to the characteristics of demand resources and are meant to foster their participation. This is the case, for instance, of the reliability-option CRMs introduced in Belgium and Italy and of the decentralized capacity obligations traded in France. For instance, in Belgium, demand resources are allowed to bid their own strike price, a key element for the settlement of the reliability option.

2.1.2.1 Baselining

Another element required to allow this kind of demand participation and the assessment of its performance is a methodology to identify a demand baseline. The latter may be used to define the firm supply of these demand resources, in conjunction with a de-rating factor. De-rating of demand resources usually depends on the self-declared or tested duration of the service that the demand aggregator can provide. The lower the duration, the lower is usually the expected contribution to scarcity conditions and, consequently, the de-rating assigned to the resource. The demand baseline is also essential to verify and quantify the compliance of these resources to their capacity commitments, by comparing the actual withdrawal with the one that would have been registered if the service had not been activated.

As for any other DR programme, several baselining approaches are possible. Some recent studies, such as the ones developed by Elia, the Belgian transmission system operator, found that the most widely adopted methods for capacity mechanisms are historical and control-group baselining. The former use historical data to estimate the expected demand in the activation period, by applying exclusion rules and rankings. For instance, high-X-of-Y methodologies focus on the last Y days of the same kind of the activation day (e.g., working days) and, within this group, select the X days with the higher load. For each settlement period, the baseline is defined by averaging the load during these X days. Historical baselines may also rely on some sort of same-day adjustment, i.e., a methodology that modifies the baseline according to the load registered during the day of activation (with expedients to avoid gaming from the demand resource to overestimate its contribution). A typical example of a historical baseline methodology with same-day adjustment used in the California ISO (CAISO) is shown in Figure 5.

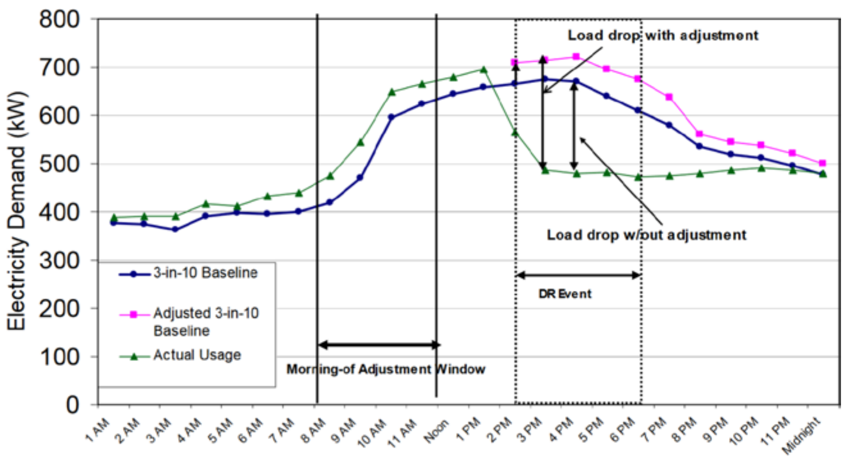

Figure 5. Historical baselining with morning-of adjustment used in CAISO

Control group baselining does not rely on historical data but estimates the load that the demand resource would have withdrawn if it was not activated based on the withdrawals of a control group of consumers. These consumers may be selected among those who are providing demand response (randomised controlled trial over a small number of active consumers), or they may be consumers with similar characteristics as those in the demand-response programme, but who do not provide DR services. This type of demand baselining is used especially in the United States.

Other baseline methodologies, which are rarely used in the framework of capacity mechanisms, are MBMA (Meter Before - Meter After, which defines the baseline based simply on the load registered before the DR service is activated), declarative baselining (the baseline is estimated directly by the DR service provider, who communicates it to the operator), and regression-based baselining (the baseline is computed through a complex formula with several parameters, as temperature and daylight, whose coefficients are defined based on historical data).

2.1.2.2 The double-remuneration (or double-benefit) problem

The supply-side explicit participation of demand resources in CRMs has a significant disadvantage that stems from the presence of a certain group of consumers both in the demand and the supply curve of the capacity market. A demand resource taking part in the capacity mechanism is remunerated for reducing its load during scarcity conditions. However, by doing so, it can also reduce its contribution to the recovery of CRM costs, if the charges are designed to allocate these costs to the withdrawals during scarcity conditions. This way, the demand resource is remunerated twice and, more importantly, its net position in the CRM could be larger than zero, i.e., it could have a net revenue from its participation.

However, what the demand resource is actually doing is just to reduce or to set to zero its load during scarcity conditions, with the final goal of avoiding the payment of CRM charges (and without benefitting from the coverage of the mechanism). Therefore, its net position in the CRM should be, at maximum, null. This point of view was clearly stated by the Italian regulator ARERA during the design of its capacity mechanism. According to its criteria, a demand resource involved in a capacity market provides a service that can only be used by itself (through the consumers it is formed by, which are also represented in the demand curve). However, the service from a demand resource cannot be provided to a third party, differently to the service provided by a generator, whose position in the CRM is of course expected to be larger than zero.

The risk of a double remuneration for demand resources depends on the design of CRM charges. Paradoxically, the double-remuneration problem has been avoided so far thanks to the inefficient cost-allocation strategies adopted in most capacity mechanisms. Volumetric charges covering a very large number of hours reduce the benefit that can be achieved by demand resources reacting during scarcity conditions. However, cost-allocation strategies based on capacity charges during scarcity conditions could increase the risk of double remuneration for demand resources. The most efficient way to deal with this problem would be, once again, to introduce fixed CRM charges (based, for instance, on historical consumption during scarcity conditions). With this approach, the demand providing DR services would pay a fixed amount of CRM costs and would offset this quantity by the revenues it receives from the capacity market. If all the elements of the CRM are properly harmonised, this combination should result in a net position close to zero, although deviations are possible.

2.2 Implicit participation

Once the capacity market is cleared and commitments assigned, there is still some space for implicit demand participation. The potential for this kind of participation mode clearly depends on the design of the charges introduced to recover the costs of the mechanism and on the signals they convey. Demand resources can basically shift their load to minimise payments derived from these charges, moving their consumption out of potential scarcity conditions.

As mentioned in section 2, the real driver of CRM costs is the contribution of each consumer to the demand for firm supply and a good proxy parameter to estimate this contribution is the historical load during scarcity conditions, with a moving average. Using historical data with a moving window certainly dilutes but does not eliminate the signal for consumers to reduce their load. If data from the last five years are used, an end-user who manages to eliminate the load during all scarcity conditions registered in the system would stop paying any CRM charge after five years. The signal could be further strengthened if charges are applied on consumption during scarcity conditions in the delivery period, although this strategy may affect cost recovery.

In real CRMs, however, implicit participation of demand resources has always been almost nonexistent, since, in the majority of cases, CRM costs are recovered through volumetric charges applied over a large number of hours, which impede an efficient reaction by the load. The United Kingdom offers a paradigmatic example. Since the introduction of the capacity market, CRM costs were recovered through a charge on electricity suppliers that was applied to the net demand (gross demand minus embedded generation) they served from 16:00 to 19:00 in the working days from November to February. This approach prompted suppliers to sign agreements with embedded generation (mainly diesel gensets) to produce in those hours, thus reducing the net demand. This kind of demand response was not efficient from an adequacy point of view, since the load reduction was taking place in hours where no scarcity was registered, and it was having a harmful environmental impact. For this reason, the cost-allocation strategy was modified in 2018 and CRM charges are now applied on the gross demand of each supplier, in order to avoid this inefficient implicit participation of demand resources. However, the inefficiency stemmed from an inefficient cost-allocation strategy, which was only partially amended by the 2018 reform.

3 CONCLUSIONS AND REGULATORY RECOMMENDATIONS

Demand resources and the flexibility they may provide to the power system are extremely valuable in achieving resource adequacy. Most capacity mechanisms in place today allow the participation of demand resources, although with different rules and different outcomes. While demand response covers a larger share of the demand for firm supply in some power systems in the United States, Europe is lagging behind and DR only accounts for less than 3% of capacity markets in the region.

This article presents a comprehensive theoretical framework for the participation of demand resources in capacity mechanisms. Its first finding is that an efficient participation of these resources depends on the definition of the demand for firm supply and the cost-allocation methodology. Regulators should define accurate methodologies that allow to compute the demand for firm supply as disaggregated as possible. The most significant advantage of this approach is that it facilitates the participation of consumers in the side of the CRM where they belong, i.e., the demand side. The regulator would define the demand for firm supply for each end-user or broader categories of them. Consumers would then be allowed to modify this value, defining the capacity they would like to be covered by the capacity mechanism and that they would be allowed to withdraw during scarcity conditions.

A disaggregated definition of the demand for firm supply also allows a more efficient cost allocation. In fact, this demand is the real driver of CRM costs; these costs, once the capacity market is cleared, should be considered as sunk costs and they will not vary if consumers reduce their demand during the delivery period below the value that was procured for them. If the demand for firm supply is disaggregated, CRM charges could be easily applied to the demand for firm supply allocated to each consumer (or broader categories), through fixed charges. Fixed charges could also be applied on a proxy basis, for instance, using historical consumption during scarcity conditions. Other approaches for cost allocation could try to foster some sort of implicit participation of demand in the capacity market, by reacting to the economic signals conveyed by charges. This goal could be achieved, for instance, by applying capacity charges during scarcity conditions, although such an approach may endanger cost recovery. CRM costs can also be recovered through simple volumetric charges applied over a large number of hours, as done today in many CRMs. However, these charges do not send any efficient signal to consumers and can be equated to a *de facto* socialisation of these costs.

If consumers are not allowed to take part in the definition of the demand for firm supply, then they can only be allowed to participate in the supply side of the capacity market, where they could sell demand-response services. This approach is actually the most widely adopted in capacity mechanisms. However, it presents several complexities, stemming from the fact that the same demand is represented twice in the capacity market, both in the supply and the demand side. This situation could result in the so-called double-remuneration problem, when a demand resource is remunerated for reducing its load during scarcity conditions, but by doing so it also reduces its contribution to the coverage of CRM costs.

The participation on the supply side also requires methodologies for the definition of a baseline. As mentioned in the article, the most widespread methodologies are historical (e.g., high-X-of-Y) and control-group baselining. In theory, demand resources should be required to provide exactly the same reliability product as other resources, since they compete in the same market. However, several regulators have defined specific products that are tailored to DR services and are meant to reduce the risk perceived by these agents and to incentivise their participation.

All these complexities for the participation of demand resources in the supply side of the capacity market, if not properly addressed, may result in significant inefficiencies in the operation of the CRM. Regulators should strike the right balance between supporting demand response and ensuring the performance of the capacity mechanism.

ACKNOWLEDGEMENTS

This article has been elaborated in the framework of the research project *Flexener*, financed by the Centre for the Development of Industrial Technology (CDTI) of the Spanish Ministry of Science and Innovation, in the context of the call *Misiones CDTI* (grant MIG-20201002).

FOR FURTHER READING

- A. Pototschnig, J. M. Glachant, L. Meeus, P. Ranci Ortigosa. "Recent energy price dynamics and market enhancements for the future energy transition". FSR Policy Brief 2022/5. 2022.

- C. Batlle, P. Mastropietro, P. Rodilla. "Redesigning residual cost allocation in electricity tariffs: a proposal to balance efficiency, equity and cost recovery". Renewable Energy, vol. 155, pp. 257266. 2020.

- P. Brito-Pereira, P. Mastropietro, P. Rodilla, L. A., Barroso, C. Batlle. "Adjusting the aim of capacity mechanisms: future-proof reliability metrics and firm supply calculations". Energy Policy, vol. 164, art. 112891. 2022.

- T. Schittekatte, I. Momber, L. Meeus. "Future-proof tariff design: recovering sunk grid costs in a world where consumers are pushing back". Energy Economics, vol. 70, pp. 484498. 2018.

- T. Brown, S. Newell, K. Spees, C. Wang. International review of demand response mechanisms in wholesale markets. Report. 2019.

- Elia, 2021. Baseline Methodology Assessment. Final report.

Biographies

Pablo Rodilla is with the Institute for Research in Technology, Comillas Pontifical University, 28015 Madrid, Spain.

Paolo Mastropietro is with the Institute for Research in Technology, Comillas Pontifical University, 28015 Madrid, Spain.

Paulo Brito-Pereira is with the Institute for Research in Technology, Comillas Pontifical University, 28015 Madrid, Spain.