MIT CEEPR

MIT Center for
Energy and Environmental
Policy Research

The Impact of Uncertainty on the Need and Design of Capacity Remuneration Mechanisms in Low-Carbon Power Systems

Fernando J. de Sisternes and
John E. Parsons

February 2016 CEEPR WP 2016-004

A Joint Center of the Department of Economics, MIT Energy Initiative and MIT Sloan School of Management.

The Impact of Uncertainty on the Need and Design of Capacity Remuneration Mechanisms in Low-Carbon Power Systems

Fernando J. de Sisternes *†, John E. Parsons ‡

Massachusetts Institute of Technology

February 18, 2016

(Revision of March 7, 2016)

Abstract

The case for or against capacity remuneration mechanisms (CRMs) is often made in a simple framework that takes the structure of the electric power system as a given and does not substantively consider uncertainty. This paper highlights the central role that uncertainty around the net load profile and net load growth plays in the case for a CRM and in determining its optimal design. Using a stylized example, the paper shows that uncertainty increases the risks investors take financing new generation so that a higher likelihood of near term profits is required. Existing alternative designs of CRMs have differing implications for how society can provide an efficient and effective signal about its demand for security of supply under uncertainty. In addition, different CRMs each present limitations in their ability to incorporate appropriate CRM design principles, which are also emphasized.

1. Introduction

In many industries, we leave it to private investors to decide how much capacity to install or hold available in order to supply a given product or service. The profit motive guides their forecasting of demand and the timely installation of needed capacity. No one expects investors to always get it right: history is replete

*MIT Energy Initiative, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E19-341, Cambridge, MA 02139-4307, USA. E-mail: ferds at mit.edu

†Argonne National Laboratory, 9700 Cass Ave, Argonne, IL 60439, USA. E-mail: ferds at anl.gov

‡MIT Center for Energy and Environmental Policy Research (CEEPR), 77 Massachusetts Avenue, E19-411, Cambridge, MA 02139-4307, USA. E-mail: jparsons at mit.edu

The views and opinions expressed in this report are those of the authors and do not necessarily reflect the position of MIT, CEEPR, the MIT Energy Initiative, or any other institution with which the researchers that participated in this study are affiliated.

with examples of over-and underinvestment. But most substitutes for private risk taking will also involve mistakes. As private investors bear much of the costs of their mistakes, they have a powerful incentive to get their investment decisions as right as they can. In the electricity industry, the so-called 'energy-only' wholesale market design operates on this principle. There are many voices, however, that advocate supplementing the energy market with a 'capacity market'—or, more generally, some form or another of a capacity remuneration mechanism (CRM)—in which regulators determine some level of needed capacity and pay a price to obtain it. CRM designs come in many different forms and a variety of labels, including capacity obligations, reliability options, and strategic reserves. There is intense debate about the need for a CRM, about which design is best, and about how to integrate CRMs into the larger market framework, whether it be a single European energy market or the US standard market design for RTOs/ISOs. This paper focuses on how issues associated with uncertainty underlie these debates and shape how well CRMs serve their purpose.

CRMs are designed to address extreme and rare events: the small number of hours each year when load strains the limits of the electric power system, causing dramatic spikes in the wholesale price and threatening system stability. At least we hope these are extreme and rare events. Or, to be more precise, capacity markets are designed to assure that these events remain rare by guaranteeing that the system has enough capacity to handle peak load at reasonable prices—a property described as *adequacy* in the literature—and safeguarding system stability (Batlle and Perez-Arriaga, 2008). CRMs were conceived before intermittent generation became a relevant factor in system operation, but intermittency has increased the concern about sufficient dispatchable generation to securely serve peak net loads.[1]

The case for CRMs is often made in a simple framework that takes the structure of the electric power system as given and does not substantively consider uncertainty. In the ideal energy-only market, the very few hours of the year when capacity is strained are expected to produce such extremely high prices such that the profits earned in these few hours cover a very large portion of the capital and fixed operation and maintenance costs incurred by all types of installed capacity. In fact, however, there is a lot of uncertainty about the frequency and severity of these scarcity events, and, more importantly, a lot of uncertainty about the value of marginal capacity in these events. When we layer uncertainty onto the problem, as we will explain, relying on this structure of profit is not a tenable way to incentivize investment. Shifting the structure of profit to one in which the same total revenue is paid for capacity across a broader number of hours provides a better, more reliable signal to investors, which lowers the cost of capacity to society. Thus, in the face of uncertainty, a well-designed capacity mechanism is preferable to an energy-only market design. At the same time, any CRM can be improved by exposing providers of capacity to price risk during

[1]The term 'net load' and 'net demand' refer to the difference between electricity demand and renewable generation.

those scarcity hours. Exposing generators to marginal price incentives during scarcity hours incentivizes the short-term optimization of capacity resources so that they are available when they are most useful and are supplied at least cost, and assures that capacity that has been paid for is actually delivered—this is what is referred to in the literature as *firmness* (Batlle and Perez-Arriaga, 2008). The mechanism should be agnostic about which technologies provide the firm capacity, so long as it is provided. The acknowledgement of uncertainty also informs us about what to expect in a good capacity mechanism over time. It must be flexible enough to reflect changing circumstances: even when the mechanism works optimally as planned, there will be times when there is excess capacity and the price of capacity is zero, and other times when a shortfall in capacity arises which requires that the price paid for capacity rises. A rigid mechanism that pays a constant price for capacity cannot adequately adjust to reflect changing conditions and will inevitably cost too much over time.

2. Why Capacity Remuneration Mechanisms?

The case for capacity markets or some other form of CRM has already been made in an extensive literature (Batlle and Perez-Arriaga, 2008)(Joskow, 2008)(Rodilla and Batlle, 2012). Here we provide a quick summary to provide a foundation for the central points of this paper. CRMs are a tool to address the 'missing money' problem that results when prices in the wholesale market do not fully reflect the scarcity value of capacity. This failure is most easily explained when there is an explicit cap on market prices during scarcity, as is common in many electricity markets. However, in practice, explicit caps have not been the main cause of missing money. Instead, when demand strains available capacity, system operators take a number of actions to manage resources without correspondingly allowing prices to reflect the full cost of those actions—whether that be voltage reductions, the dispatch of certain generators out of market, and other emergency protocols including, in the extreme case, the use of rolling blackouts. In all of these cases, the market price seen by generators is often far below the true marginal value of a unit of capacity, so that some analysts speak of a *de facto* price cap (Cramton and Stoft, 2006)(Joskow, 2008). We will use the term price cap in this sense, keeping in mind that sometimes there is no explicit cap and that even the *de facto* cap is not set at a specific level. In the absence of CRMs, price caps mean that market prices do not reflect the actual value of the firm capacity installed and do not provide sufficient incentives to support the efficient quantity and mix of generation capacity. CRMs replace the extremely large payments in these rare events with a much smaller payment made to capacity across all hours.[2] The smaller payment can be thought of as a kind of insurance premium paid to avoid the very high prices that would otherwise be necessary to accurately reflect the value of capacity during peak load events (Vazquez and Perez-Arriaga, 2002).

[2]In expectation, the sum of these frequent smaller payments should be equal to the sum of the infrequent larger scarcity rents.

CRMs of one form or another are now a reality in many liberalized electricity markets in the United States, Latin America and Europe, although the path towards their introduction has varied across markets due to different historical developments that drew attention to the issue of capacity adequacy (Perez-Arriaga, 2013)(Sioshansi, 2008)(Batlle and Rodilla, 2010). The liberalization of the electricity generation sector that began in the early 1980s proceeded in most countries against a backdrop of excess capacity, so that the energy-only market design functioned satisfactorily, its defects being purely hypothetical. In some cases, subsequent GDP growth accompanied by growth in peak electricity demand put capacity-adequacy back onto the agenda. In other cases, climate policies spurred investments in and dispatch of renewable resources, altering the pre-existing equilibrium of quantity and mix of generation capacity installed and undermining the profitability of existing thermal units. In Europe, the ongoing economic crisis since 2008 sharpened the drop in utilization of these thermal units and increased the number of threatened and actual closures—either temporary or definitive—.[3] At the same time, growing reliance on intermittent renewables has also increased the likelihood of events during which net load presses against dispatchable capacity. This has triggered occasional price spikes that have prompted some regulators to introduce price caps to reduce the possibility of exercising market power, and it has put the need for capacity markets back onto the agenda (Linklaters, 2014)(ACER, 2013)(FERC, 2013).

3. Uncertainty

The simplest case for a capacity market is made, naturally enough, with a simplified picture of demand and supply. In particular, given a known distribution of demand and output from renewables, the net load duration curve can be mapped.[4] Next, given a known capacity of dispatchable generation resources, the

[3]Closures are motivated by two short-term effects that result in a reduction of the revenues earned by generation assets: a reduced utilization of the asset, and declines in electricity prices. The severity of these two effects is experienced differently by different types of generation: peaker generators will be exposed to both price declines and lower utilization rates, as peaker generators are displaced by low-variable-cost renewable generators or by reductions in net demand; whereas baseload generators will be exposed to price declines too, but their exposure to lower utilization will be smaller since they are less often displaced from the merit order as a result of their low variable cost. As a consequence of the extra variability of renewable generators, the net load duration curve (NLDC) resulting from renewable deployments will have a lower minimum net load level and a steeper distribution of peak net load events than the load duration curve (LDC). In the long-run, the new capacity mix in equilibrium resulting from such NLDC will have less baseload capacity and more peaker capacity than the original mix without renewables since baseload generators will be running for a lower number of hours. Here a paradoxical situation appears: in the short-run, peaker generators are the most exposed technology to lower prices and lower utilization; yet, in the long-run equilibrium, once part of the existing baseload capacity exits the market, those same peaker plants become more economically efficient than baseload plants, and might also become essential to the system's security of supply because of their greater ability to provide operating reserves.

[4]The net load duration curve is the series of net load data points across one year, ordered in a decreasing net load fashion. It can be thought of as the cumulative distribution function of the net load rotated 90 degrees clockwise.

frequency of scarcity conditions can be determined. Under these assumptions, it is possible to describe the operation of an energy-only market and the frequency of scarcity periods in which demand exceeds available supply and market prices spike. With the frequency and severity of scarcity prices known *ex ante*, investors could make fully informed investment decisions about the capacity of different types of plants to serve different segments of the net load duration curve. An energy-only market would thus allow perfect cost-recovery of the optimal generation quantity and mix, provided that market prices are allowed to rise during scarcity periods. In contrast, layering a price cap onto this market, whether *de jure* or *de facto*, undermines cost-recovery for all types of generation in the mix. In this case, introducing a CRM restores the possibility of cost-recovery, and once again incentivizes an optimal generation mix. Given the known net load duration curve, the precise amount of capacity required is easily calculated, as is the size of the capacity payment required to restore the 'missing money'.

While this simple case is useful for introducing and explaining the basic problem, it is important to look beyond this stylized example and appreciate how uncertainty complicates the problem. Without uncertainty, the differences between existing market designs are quite trivial. Without uncertainty, choosing the right capacity portfolio among different technologies is a straightforward decision. Moreover, the revenue requirement for different units of capacity is also clear, and thus it is straightforward to identify alternative mechanisms to provide that revenue. An energy-only wholesale price without a price cap is one of the many alternatives that satisfy the revenue sufficiency requirement, as is a simple capacity payment scheme. There are many alternative CRMs that all satisfy the revenue sufficiency requirement and, ignoring uncertainty, could lead to the same efficient mix of generating resources. The real differences between these alternative CRMs only become clear when we introduce uncertainty into the picture. And only with uncertainty does the real motivation for exposing generators to market prices enter into the discussion and lead to the preference of one CRM over another.

There are a number of different sources of uncertainty that affect investments in electricity generating capacity and the design of CRMs. First, there is uncertainty about demand. The realized load duration curve in any given year depends upon a variety of drivers, including variations in weather that change the demand for air conditioning or heating and the health of the overall economy, which drives demand for energy. This uncertainty operates at different time scales: in any given year, there is uncertainty around a known baseline demand level, and looking forward a number of years, there is additional uncertainty about where that baseline will be, depending upon economic growth, upon changing technologies related to energy use and therefore on changing patterns of demand for electricity, and—in the future perhaps increasingly—also depending upon changing weather patterns. Second, there is uncertainty about the resource availability of intermittent generation, affecting the production of renewable energy at different times. This uncertainty,

too, operates at different time scales. In any given year, there is uncertainty around a known baseline determined by the typical weather patterns affecting wind and solar energy output. Third, there is uncertainty about realized dispatchable capacity, given whatever installed base is in place. Generating plants sometimes break down unexpectedly, and transmission lines come down or get temporarily derated.

All of the above factors are exogenous uncertainties, at least as far as the electricity policy and market design are concerned. We call them exogenous uncertainties because they affect the outcome from the market design, but the market design does not directly affect their realizations—at least not first order.

Other factors exist that are variable and may seem like uncertainties from the viewpoint of policy makers, but that are at least in part affected by the market design. For example, the very purpose of CRMs is to incentivize the installation of capacity. Management of dispatchable plants involves choices in the timing of maintenance activities, which shifts available capacity from one time of year to another, or which may reduce or increase overall future downtimes, and different CRM designs change the incentives for these decisions. Policy makers choose on behalf of society the incentives for installation of new renewable capacity, and the operation of a CRM may interact with decision-making on renewables since the CRM would provide a stronger incentive to renewable technologies that are firmer than others. Insofar as the market design affects the realization of these factors, we call these endogenous uncertainties. Another important source of uncertainty that needs to be included in the policy discussion around CRMs is society's willingness to pay for security of supply. What is an acceptable loss of load expectation (LOLE)? Or, to put it differently, what is the value of lost load (VOLL)? Reliability standards involving the LOLE and VOLL codify what is essentially a social choice. Different CRM designs shape the ongoing public policy discussion and ultimately the choices made.

While the simple case for capacity markets is made assuming away uncertainty, the actual design of CRMs takes some of the foregoing sources of uncertainty into account, especially the exogenous factors operating at the time scale of a given year. The design of CRMs typically relies on probabilistic models that reflect different weather conditions, load, renewable resource availability, and transmission availability, to analyze the effect of different planning reserve margins on the reliability of the system. Model outcomes are subsequently used to determine the planning reserve margin that minimizes expected reliability-related costs and the economic incentives necessary to secure those margins, as it is done for instance in (Brattle, 2013).

Arguments in favor of certain CRM designs over other CRM designs are based in part on how some of the endogenous uncertainties should be handled. Reliability options, for example, are designed to expose

generators with capacity obligations to wholesale price risk when the security of supply is threatened. In that way, the providers of capacity are incentivized to anticipate supply shortfalls and schedule their maintenance to be available at the most critical times, and similarly to carefully weigh the value of increased maintenance that brings increased availability overall. In this fashion, reliability options are intended to purchase capacity for society when it is most needed and to make that capacity available at least cost.

Making uncertainty explicit can also sharpen the discussion since differences of opinion about the exogenous uncertainties often underlies debates about capacity prices and alternative CRM designs. While academic debates may focus on idealized situations, debate in the body politic is always concerned with the specific historical situation at hand. Some parties objecting to a capacity market are concerned that the price is too high in light of their assessment of the likelihoods, while others who assess the dangers of insufficient capacity differently may be anxious that even these prices may not induce enough new investments. Inherited reliability standards and legacy methodologies for acceptable LOLE and estimates for VOLL can make it seem as if society's willingness is a settled matter, but it never is. The re-emergence of capacity adequacy as a policy issue always entails new debates on this issue and a reevaluation of legacy methodologies.

Differences of opinion about the uncertainties are often surprisingly intractable, given the many exogenous factors that should be amenable to objective measurement. It is important to keep in mind, however, that the issue of capacity adequacy is not really a debate about the average events, but rather a debate about extreme or tail events. By their nature, these events are rare. Therefore, making a reliable inference from the historical record of the probability of such extreme events is inherently difficult. Moreover, the system is always undergoing change, so only the most recent history is really relevant, further limiting what we can reliably infer. This is true for the exogenous uncertainties operating within a short time scale like a year: there can be significant difference of opinion about the danger of extreme tail weather events and the consequent extreme tail demand. It holds even more for uncertainties operating at the longer time scale relevant for investment decisions. And at that time scale, the endogenous factors also start to be relevant and create even more differences of opinion: policy makers with different visions for the pace of additional renewable capacity are going to have different views about the risks to capacity adequacy. Even if policy makers were able to project a clear vision for future renewable goals, investors considering the profitability of new back-up plants would likely be very circumspect. A key to evaluating alternative CRM designs is how the different designs interact with these differences of opinions, and how that shapes the market outcome.

4. A Model to Illustrate One Impact of Uncertainty on Security of Supply

Under ideal conditions, energy-only markets should produce a price signal that supports the optimal, minimum cost portfolio of generation investments (i.e., a price signal that induces an efficient level of investment in generating capacity and its efficient operation). These ideal conditions can be summarized as: 1) the functions representing generators' cost structures and demand utility from electricity are convex (i.e., there are no discontinuous jumps from startup costs, etc.); 2) there are no economies of scale; 3) investments in capacity are not lumpy; and 4) the market is perfectly competitive with perfect information (Schweppe et al., 1987)(Perez-Arriaga and Meseguer, 1997). Wholesale prices in energy-only markets reflect the marginal cost of electricity supply; that is, during non-scarcity periods, the marginal operating cost of the most expensive generating unit; and during scarcity periods, demand's marginal utility of consumption (i.e., the VOLL).[5] Optimal prices derived in an energy-only market should therefore be allowed to rise high enough to clear the market when generation capacity is fully utilized, permitting all types of generation in the optimal generation mix to earn sufficient infra-marginal rents to perfectly recover fixed costs. Failing to do so leads to situations of underinvestment that increase the probability of system operators resorting to rolling blackouts, or even an eventual collapse of the network.

Many studies have identified causes of market failure that violate some of the theoretical conditions that support the efficiency of energy-only markets, producing deviations from the optimal capacity mix (Perez-Arriaga and Meseguer, 1997)(Hogan, 2005)(Joskow, 2008). One important issue that has become salient as policy-makers take action on reducing the carbon footprint of power systems is the common practice to provide incentives to increase the deployment of renewable generation capacity. Decisions about such incentives are rarely, if ever, communicated to market participants early enough to anticipate the resulting market changes and adjust their investment decisions in an optimal manner. This could constitute a case of information asymmetry, leading to a sub-optimal quantity and mix of generation capacity that is not 'well-adapted' to the new total renewable capacity deployed (de Sisternes et al., 2015).

Starting with a situation of perfect information, we layer up the two uncertainty considerations at issue, and explain the limitations of energy-only markets in addressing each of them, underscoring the importance of CRMs in filling the identified gaps.

[5]Without loss of generality, our argument assumes away the locational effects of network congestion and losses on prices, which would be reflected in locational marginal prices.

4.1 Perfect information: the certainty case

Under idealized perfect information conditions, the distribution of demand and renewable output are known *ex ante*, demand growth and renewable deployment pace are also known *ex ante*, and so is the frequency of scarcity conditions and price spikes. Investors can make fully informed investment decisions, and there would be no controversy among consumers and the regulator when price spikes appear. An energy-only market would thus allow perfect cost-recovery of the optimal generation quantity and mix. Under these conditions, a regulator may choose to layer a price cap onto the energy-only market that reduces the opportunities to exercise market power, while also introducing a CRM to restore optimality.

The certainty case is depicted in Exhibit A using a single-technology model. In this model, the regulator decides on a particular standard of security of supply—in this example represented by the number of hours with non-served energy—the LOLE. The LOLE is then used to determine a scarcity price (VOLL) that elicits the amount of generation capacity necessary for the system to comply with that required LOLE. Scarcity rents earned by the installed capacity are sufficient to perfectly recover capital costs,[6] resulting in the economic equilibrium represented by the optimal capacity-demand ratio[7] depicted in Exhibit B.

4.2 Uncertain net demand distribution: the stationary case

Intermittent energy resources—including wind and solar power—increase the volatility of energy supply as well as the associated market clearing prices, creating uncertainty around when the system will experience scarcity. It takes time to fully understand the underlying process characterizing demand uncertainty, and it is the limited knowledge about this process, built upon historical experience, that informs investment decisions.

While this learning quickly resolves uncertainty about the average level of future demand, uncertainty about the tail and the likelihood of extreme events is much slower to be resolved.[8] Therefore, disagreements on the tails of the demand distribution will be more frequent than disagreements on the average.

[6] For a full exposition of cost-recovery under perfect information refer for instance to (Joskow, 2008), (Rodilla and Batlle, 2012) or (Perez-Arriaga and Meseguer, 1997).

[7] The capacity-demand ratio is simply the quotient between the total capacity installed and the maximum demand in the system.

[8] One common characteristic of stochastic processes is that one learns faster about the average of the underlying probability distribution characterizing the process than about the tails of that distribution (i.e., it takes more observations to obtain an estimate for the variance of the underlying probability distribution at a given level of significance than it takes to obtain an estimate for the mean at the same level of significance).

The different cost structures of the different types of generation make them unevenly affected by the tails of the distribution (Joskow 2008): baseload generators earn most of their profits in the central part of the demand distribution, while peaker generators earn most of their profits during the tails. Therefore, investments in peaker generation are more controversial than investments in baseload generation, as there is more disagreement about the parameters characterizing the tail of the distribution.

Some studies have pointed at the combined volatility of demand and renewable generation as one possible cause for underinvestment in generation capacity. However, a rational investor with perfect information about how the net demand is distributed—and therefore perfect information on its volatility—would have little trouble estimating its expected earnings over the life of the asset. Therefore, it is not the volatility *per se* that is problematic, but the uncertainty on the underlying parameters and processes characterizing that volatility (i.e., uncertainty about the variance of demand and renewable output, uncertainty about the mean and variance of demand growth, and uncertainty about the process characterizing renewable capacity growth).

In an energy-only system, disagreements among regulators, consumers and investors on the parameters of the underlying distribution are not confronted except in the event of price spikes. It is possible for consumers to assign a low probability to tail events, while investors in peak generation place a high probability. This is a recipe for conflict *ex post*. *Ex ante* disagreements about the tail translate into sharp *ex post* disagreements about price spikes, and their underlying causes. Consumers or regulators who do not expect a large tail are quick to identify alternative nefarious causes for spikes, subjecting investors in peak generation to ex post confiscation of expected returns in the form of price caps. The CRM forces the resolution of this discussion to occur *ex ante*. This is good for investors in peaking capacity, because they can be more confident in their compensation and therefore more reliably invest. It is also good for consumers, by ensuring that an adequate level of firm capacity is in place, reducing their exposure to the high costs of tail events. Yet it also forces a public debate *ex ante* about a subject on which the public finds it difficult to be informed—tail events—and on which great disagreement is common. Therefore, a well-designed CRM is likely to provide more assurance to investors and consumers and therefore a lower average cost for whatever capacity is to be procured.

4.3 Uncertain demand growth: the dynamic case

Reliability analyses determine the optimal necessary capacity-reserve margins in the system—the required amount of generation capacity—simulating for a given year the reserve margin in the system under many scenarios that reflect various contingencies (e.g., peak demand, renewable output, weather conditions etc. See, for example, (Brattle, 2013)). Additional uncertainty is introduced by uncertain growth in demand

and uncertain growth in renewable capacity—which together lead to uncertain growth in net demand. This quickly expands the margin required as we look out beyond the current year.

As indicated above, Exhibit A shows a stylized model that determines the level of generation capacity investment required to meet a pre-specified reliability standard, provided by the LOLE. This stylized model assumes that the capital cost annuity recovered from operating the generation capacity during the year of the simulation will also be recovered during the subsequent years spanning the life of the asset. In reality, however, this will not be the case if net demand decreases during certain periods as a result of economic cycles or regulatory interventions that unexpectedly bring new out-of-market generation capacity. Therefore, there is a mismatch between the reliability outcomes derived from the stationary treatment in reliability studies of the growth in net demand and those found in real-world markets, where net demand experiences dynamic growth. With dynamic uncertainty the reliability in the system varies throughout time, depending on the legacy generation capacity at a given point in time as well as on new plants whose expected operating profits at the time of investment—taking into account the capacity-demand ratio at the time of investment and the dynamic evolution of the net demand— were equal to their investment cost.

This dynamic uncertainty interacts with each market design. In an idealized energy-only market without any 'missing money' problem, there will be times when generation capacity is well-adapted to a particular net demand profile, but subsequent events produce a lower profile of demand than had been anticipated (such as an economic recession, new out-of-market renewable capacity, or newly available technologies to improve energy efficiency). In that case, there will be an excess of capacity, a lower profile of wholesale prices, and installed generation capacity will not earn enough revenue to cover its sunk capital cost. This will be offset, in expectation, by times when subsequent events produce a higher profile of demand than had been anticipated, there is a shortfall of capacity, and the profile of prices is higher so that capacity is earning revenues far above what is necessary to recover sunk costs. However, in an energy-only market with the 'missing money' problem, the dynamic uncertainty interacts in an especially harmful way. When demand is growing unexpectedly quickly, this will increase the number of scarcity hours and therefore increase the scale of the 'missing money' problem. Unfortunately, this is exactly when additional capacity is more urgently needed. Thus, the disruption to the signal is greatest when the signal needs to be the strongest. A capacity market, in restoring the 'missing money', whether it is a large or a small amount, helps to send the appropriate signal about the need for new capacity. In a capacity market, times with unexpectedly low demand will lead to falling capacity prices. Capacity that had been procured in forward markets will now be available below those legacy forward terms, and load will regret having locked in the forward price. In contrast, times with unexpectedly high demand will lead to rising capacity prices. Additional capacity beyond the amount already procured in forward markets will now only be available above those legacy

forward terms and load will be happy that it had locked in the forward price on some capacity earlier. The different market designs—whether energy only, capacity market, or some other CRM— and the duration of the forward contracts resulting from each of them, shape the degree of different parties' exposure to these *ex post* gains and losses. The different incidence of these ex post gains and losses feed back as incentives to the parties that will bear them and ultimately shape investors' decision on when to install new capacity.

In a similar manner as with the treatment of uncertainty around net demand distribution, investors learn about the distribution of net demand growth based on historical experience. Characterizing the out-of-market provision of subsidies to certain technologies that change the shape of net demand seems to be an almost impossible task, however. Using historical information, an investor will protect investments against the possibility of declines in net demand. This action can lead to the withholding of investments in new capacity until periods with greater scarcity rents that protect the asset against the downside of potential negative outcomes. In economic equilibrium, this behavior would translate into a capacity-demand ratio lower than the one obtained under perfect information, if the same VOLL is used in both cases (Exhibit C).

This effect has important implications for the electricity system's security of supply, as the average equilibrium amount of non-served energy with net demand growth uncertainty is greater than the amount obtained with perfect information on how net demand growth is distributed.[9]

5. Existing Solutions to, and Implications of, the Uncertainty Dimension

Different existing implementations of CRMs and procedures aimed at addressing the problem of security of supply imply different treatments of the uncertainty issues discussed in this paper and, in practice, will likely lead to different equilibria.

5.1 Energy-only markets

Energy-only markets rely on prices rising high enough during scarcity periods to create a positive profit expectation at the optimal amount of installed capacity (Hogan, 2005). Traditionally, maintaining prices that are sufficiently high has encountered numerous political barriers, as price spikes sometimes are perceived—mistakenly or not—as a reflection of the exercise of market power and persistent supply scarcity is often considered by policy makers a crisis that necessitates various interventions. If prices are capped below the

[9]It is important to emphasize that this result follows with no risk aversion on the side of the investor, showing how the investor's risk-neutral reaction in the face of uncertainty is to withhold investments until the volume of energy scarcity is sufficient to protect him against the downside of periods with low demand and excess of capacity.

level necessary to achieve the required reliability criteria, an energy-only market alone will not be able to provide the incentives necessary to attain the optimal level of investment. Arguably, the design of energy-only markets has also evolved over time in ways that have helped to mitigate the missing money problem. How satisfactory this evolution has been and how far it could go is a matter of debate, as will be shown below.

The ERCOT system in the U.S. is loosely identified as one of the very few real power systems implementing energy-only markets. In effect, however, ERCOT applies a bidding cap together with an operating reserve demand curve that reflects the value of available reserve capacity (Hogan, 2013). The reserve demand curve creates a real-time reserve price that is added to the locational marginal pricing-based energy price and is allowed to reach the VOLL during periods where reserves do not meet a minimum established level. What the operating reserve demand curve in ERCOT means in practice is that the VOLL is triggered during periods of capacity reserve scarcity, as opposed to periods of scarcity in energy supply. This mechanism partially addresses the uncertainty around extreme events by linking operating profits to the provision of a reserve margin. This is an important example of improving the energy-only market by making prices better reflect the full scarcity value and thereby reducing the missing money problem. However, the impact is limited, and much of the problem remains. In particular, this only partially addresses the very short-term uncertainty, but does little to mitigate the uncertainty on net demand growth.

Moreover, determining the VOLL—or the reserve demand curve—that achieves the optimal capacity-demand ratio under uncertainty in net demand growth or under other real-world uncertainties can prove to be an extremely difficult task. Exhibit D shows how a sufficiently high VOLL can drive investments close to the optimal capacity-demand ratio. However, since the earnings of capacity investments rely solely on the uncertain frequency of extreme events, the VOLL calculated can easily elicit an amount of generation capacity that deviates from optimal conditions. This deviation can motivate *ex post* disagreements about price spikes or the profitability of some assets. Energy-only markets thus impede reaching an *ex ante* agreement on any adequacy or capacity margins, and there is no guarantee that the desired level of security of supply will be achieved.

Energy-only markets can also be coupled with that allow load to hedge itself against potentially high scarcity prices, as advocated by Harvey and Hogan (Harvey and Hogan, 2001) and as happens for instance in ERCOT and in the National Electricity Market (NEM) in Australia. With hedges in place, it is more feasible politically to lift the price cap on scarcity events and reduce the missing money problem. However, in emergency situations where the system operator has to implement rolling blackouts, all customers are affected indiscriminately by the rolling blackouts regardless of their long-term contracting. Bilateral contracting,

therefore, can give rise to a collective action problem—also referred to as free-riding problem—, whereby participants engaging in these contracts benefit other energy consumers by increasing the generation capacity in the system, reducing their probability of emergency load curtailment. This is likely to lead to suboptimal contracting and therefore bilateral contracts are unlikely to fully address security of supply challenges.

5.2 Capacity Payments

Capacity payments are a price-based CRM whereby a fixed price is paid to generators for their available firm capacity provided to the system. The regulator determines the price to be paid for firm capacity, and the duration of the payments, but does not establish the total quantity of generation capacity entitled to receive this payment. The regulator also determines the fraction of firm capacity from each type of generation technology, either *ex ante* according to historical records on each technology's availability during scarcity conditions, or *ex post* according to each generator's actual performance. Some countries exclude some types of firm capacity (e.g., old capacity, curtailable demand, or other technologies) from receiving these capacity payments, which constitutes a subsidy to the non-excluded types of capacity. Discrimination is not desirable, as it precludes already installed firm capacity from receiving the rents confiscated by the *de facto* or *de jure* price cap and shifts the quantity and mix of capacity in equilibrium away from the optimal conditions.

Capacity payments address the *ex ante* disagreement on the distribution of demand and net demand growth. This is implicit in a capacity price that reflects the value that consumers place on firm capacity, which depends on the VOLL and the assumed frequency of extreme events. Capacity payments shift part of the operating profits from the energy market to the CRM, reducing the negative impact of demand declines on the operating profits earned by the generation capacity in the program—the actual reduction of uncertainty achieved will depend on the duration of the payment and how frequent the payments are updated. The capacity payment, however, is unilaterally determined by the regulator, which impedes investors from revealing the real price level that would make them break even at the optimal capacity-demand ratio. Therefore, capacity payments can lead to situations of over- or under-investment, as there is a practical difficulty in estimating a price for firm capacity and duration for the payments that elicit the optimal level of investment. Exhibit E illustrates the effect of an increased contract duration on moving the equilibrium capacity-demand ratio closer to the optimal level.

In addition, capacity payments must incorporate incentives for capacity to be actually available during scarcity conditions and must have verifiable and efficient ways to certify the reliability of units that are not producing, but that receive the capacity payment.

5.3 Capacity Markets

Capacity markets are quantity-based mechanisms whereby a certain amount of firm capacity (or another capacity product typically covering the maximum demand plus some reserve margin) has to be contracted between load serving entities and capacity suppliers. The two major forms of capacity markets are capacity obligations and capacity auctions, differing on the degree of centralization of the procurement of firm capacity. Capacity obligations establish that load serving entities purchase their own generation or engage in bilateral contracts for firm capacity, for a total volume corresponding to a reliability margin in addition to their expected consumption. Conversely, capacity auctions centralize the procurement process of firm capacity, and the price for the capacity supplied is determined by the settlement between a capacity demand curve that reflects the required reserve margin to maintain the desired LOLE—or some other reliability metric—and the quantity-price bids presented by capacity suppliers. Blended versions of these two forms also exist.

Capacity markets decide *ex ante* the amount of capacity to be contracted, which sets a de-facto agreement on the capacity requirements for the load serving entities. In that sense, capacity markets can guarantee that the targeted amount of capacity is realized (Exhibit F), and that the price resulting from the settlements is fair to suppliers and the most efficient way to procure the specified quantity of capacity to consumers.[10] Yet, there are a number of design parameters that can critically affect the way capacity markets address uncertainty and allow generation capacity to recover costs. The first is the duration of the contract, which should be long enough to reduce some of the uncertainty on demand growth and energy policy, and short enough to foster competition between existing and new firm capacity. An important constraint on the duration of the capacity commitment is the reliability of the obligation to make the payment even in times when wholesale prices are declining or have long stagnated so that the capacity price is deep out of the money. Someone has to pay up, and the question is who and how can that obligation be enforced, and whether it is efficient to enforce it at all. The second one is the lead-time—the time between the auction and the actual physical delivery— which should be sufficient to allow new entrants to enter the market. A satisfactory design of a capacity market should be able to strike a balance between these parameters such that demand for capacity is fulfilled at minimum cost when it is needed, and that all investments recover costs.

Capacity markets should also establish a penalty for non-performance, and this performance should be monitored daily as an incentive to enhance the firmness of the capacity contracted. This is a problem that has been brought into focus in capacity markets in the U.S. (Bhagwat et al., 2016a), and it will be particularly relevant as power systems decarbonize and the operation margin seen by all generation technologies in the

[10]Note that in excessively concentrated markets, suppliers might have space to exercise market power in capacity markets, artificially increasing the capacity prices resulting from capacity auctions.

system during non-scarcity periods is reduced.[11]

5.4 Reliability Options

Reliability options are forward quantity-based mechanisms whereby the regulator requires load serving entities or the system operator to buy a volume of reliability contracts from suppliers on behalf of demand. Contracts are settled in a centralized auction where new and existing generation, as well as import capacity and demand resources, can issue bids to supply the reliability service. The contracts auctioned consist of a financial call option with a strike price and a penalty for non-delivery (Vazquez and Perez-Arriaga, 2002). The premium required by the suppliers bidding in the auction is independently decided by each of them and should reflect the foregone profits when the wholesale price is above the strike level as well as the probability of not being available to provide the capacity committed in the auction times the penalty for non-delivery. This premium is paid to suppliers in exchange for protecting demand against price spikes (Vazquez and Perez-Arriaga, 2002).

Reliability options have successfully been introduced in Colombia, and they are expected to be implemented in Italy. The implementation details of reliability options across different regions can vary, and the penalty can either be unbounded or limited to the total revenues obtained by the generator. Either way, the more reliable generators are, the more competitive they are in this market (all else equal).

The pre-specification of the volume of reliability contracts to be settled in the auction reflects an implicit ex-ante agreement on the distribution of the tails of the demand curve and addresses the uncertainty problem around the annual frequency of extreme events. Additionally, a lead-time of the contracts comparable to the construction time of new units avoids some of the market power issues found in earlier versions of capacity markets, allowing enough time for new capacity to compete in these markets. More important, the duration of the contracts (typically between 1-3 years) plays a critical role addressing long-term investment risks and reducing expected reliability costs, as they partly reduce the uncertainty around demand growth and energy policy. Longer duration contracts might be introduced if that can further reduce expected reliability costs, and so long as the obligation to pay when prices are low can be enforced.

Reliability options can be considered superior to the original design of capacity markets, which did not place a strong emphasis on non-compliance—not being firm (Vazquez and Perez-Arriaga, 2002). Yet, current designs of capacity markets have introduced penalties for non-compliance in a similar way as reliability

[11] The reduction of the operation margins during non-scarcity periods will be reflected on a larger incidence of zero hourly energy prices. On average, however, the average operation margins should be equal to the capital costs in equilibrium, accounting also for the hourly price spikes during scarcity periods.

options, such that the total remuneration received is linked to the generators' actual performance during scarcity conditions (Cramton and Stoft, 2005)(Joskow, 2008).

5.5 Strategic Reserves

Strategic reserves are quantity-based mechanisms that withdraw a pre-specified volume of generation capacity from participation in the energy market, instead letting the system operator call upon this reserve capacity during periods when demand is close to creating scarcity situations. The system operator determines the total amount of reserves to be procured in a centralized public auction, the types of generation capacity that can participate in this auction, the price at which reserves are offered in the market—typically higher than their variable cost, and lower than the VOLL—, and the parameters under which reserves are activated. In any of its forms, the system operator offers electricity from the strategic reserves to the market at a price above their variable cost and below the VOLL, artificially capping the price—or introducing a price deadband—which ultimately can create missing money.

In cases where auctions have been used to implement a strategic reserve, the auctions have been restricted to limited types of generation and, given their low opportunity cost of being withdrawn from the market, old and expensive plants at risk of being mothballed tend to be the most competitive type of generation bidding into such an auction (Bhagwat et al., 2016b). This regulatory intervention does not contribute to the main objective of CRMs, which should help reduce the uncertainty around extreme events in net demand or net demand growth to all technology options capable of providing firm capacity, in order to move the capacity supplied in equilibrium towards the optimal quantity and mix. Instead, strategic reserves distort the price signals and create an artificial asymmetry in the uncertainty seen by different types of generation—strategic reserves do not affect the uncertainty faced by generators not included in the reserve—, crowding out marked-based investments and moving generation capacity away from the optimal quantity and mix.

In practice, the main motivation for introducing this mechanism has been to avoid the mothballing of old and expensive plants. Yet, it is not a long-term efficient solution to the challenge of eliciting investments in new firm capacity. Moreover, the parameters characterizing the operation of strategic reserves increase the regulatory uncertainty perceived by investors. This uncertainty can undermine the amount of capacity installed in the long-run equilibrium, potentially necessitating a more costly CRM to address the shortfall. No RTO in North America has implemented this option to date.

6. Recommendations and Design Principles

A capacity remuneration mechanism can be a useful element of market design. It can offer a way for society to provide a rational signal about its short-and long-term demand for security of supply and to commit to paying for that supply on reasonable terms. The energy-only market does not provide a rational and reliable substitute. In the face of uncertainty about demand and renewable energy output, the energy-only market's reliance on rare events to supply a large amount of quasi rents invites *ex post* opportunism, whereas the capacity market forces an *ex ante* commitment that is more reliable to investors and therefore cheaper for consumers. Energy-only markets also present limitations addressing the long-term uncertainty about net demand growth, either through using bilateral contracting or increasing the VOLL. Moreover, in many jurisdictions the energy-only market has not yet evolved adequately—mainly due to *de facto* price caps—, meaning that a significant missing money problem remains even without any ex post opportunism.

Suppliers of capacity need to be given incentives to perform, and to perform efficiently. Capacity should face penalties for non-performance that produce incentives for short-run optimization. Improving the energy-only prices during scarcity events, and using these prices as an incentive to committed capacity, is an important step. Strategic reserves dampen the price signal during scarcity events and do not provide efficient incentives for plants participating in the market to be available during such events.

Capacity prices and contract durations should adjust to and reflect the uncertain evolution of demand and the relative surplus or deficit of capacity. Capacity payments and capacity markets can naturally and efficiently adapt to reflect these two conditions. Conversely, strategic reserves degenerate this important signal, as they could reduce electricity prices and undermine the economic viability of generators participating in the wholesale market. Moreover, strategic reserves cannot address the uncertainty in demand growth seen by investors in capacity participating in the wholesale market, sustaining boom and bust investment cycles.

Capacity mechanisms should be technology neutral, and accept the participation of any element in the system that can provide firm capacity (i.e., thermal and renewable generators, demand programs, energy storage, etc.). Failure to do so would introduce a hidden subsidy to those technologies that do qualify for the CRM program, and result in a more expensive generation mix for the consumer. In addition, performance penalties will ensure that resources are only compensated for the firmness they actually provide, which creates a level playing field between technologies and allows investors to commit the level of capacity they think they can really deliver.

Lastly, regions aiming at integrating their energy markets should leave it to individual regions to establish their respective reliability requirements, as these requirements largely depend on the particular characteristics

of the individual electric power systems. However, it would be desirable to have a harmonized denomination of the different parameters defining the CRM and scarcity events to facilitate an efficient exercise of cross-border capacity contracts.

Acknowledgements

The authors would like to thank Michael A. Mehling, Jesse D. Jenkins and Christian Stoll for their valuable comments.

References

ACER (2013). Capacity remuneration mechanisms and the internal market for electricity. Technical report, The Agency for the Cooperation of Energy Regulators.

Batlle, C. and Perez-Arriaga, I. J. (2008). Design criteria for implementing a capacity mechanism in deregulated electricity markets. special issue on capacity mechanisms in imperfect electricity markets. *Utilities Policy*, 16(3):184–193.

Batlle, C. and Rodilla, P. (2010). A critical assessment of the different approaches aimed to secure electricity generation supply. *Energy Policy*, 38:7169–7179.

Bhagwat, P. C., de Vries, L. J., and Hobbs, B. F. (2016a). Expert survey on capacity markets in the us: Lessons for the eu. *Utilities Policy*, 38:11–17.

Bhagwat, P. C., Richstein, J. C., Chappin, E. J. L., and de Vries, L. J. (2016b). The effectiveness of a strategic reserve in the presence of a high portfolio share of renewable energy sources. *Utilities Policy*, 39:13–28.

Brattle (2013). Resource adequacy requirements: Reliability and economic implications. Technical report, The Brattle Group.

Cramton, P. and Stoft, S. (2005). A capacity market that makes sense. *The Electricity Journal*, 18(7):43–54.

Cramton, P. and Stoft, S. (2006). The convergence of market designs for adequate generating capacity with special attention to the caiso's resource adequacy problem. A white paper for the electricity oversight board.

de Sisternes, F. J., Webster, M. D., and Perez-Arriaga, I. J. (2015). The impact of bidding rules on electricity markets with intermittent renewables. *IEEE Transactions on Power Systems*, 30(3):1603–1613.

FERC (2013). Centralized capacity market design elements. Commission staff report, U.S. Federal Energy Regulatory Commission.

Harvey, S. M. and Hogan, W. W. (2001). Market power and withholding. Technical report.

Hogan, W. W. (2005). On an "energy-only" electricity market design for resource adequacy. Technical report, Center for Business and Government, John F. Kennedy School of Government, Harvard University, Cambridge, MA.

Hogan, W. W. (2013). Electricity scarcity pricing through operating reserves. *Economics of Energy & Environmental Policy*, 2(2).

Joskow, P. L. (2008). Capacity payments in imperfect electricity markets: Need and design. *Utilities Policy*, 16:159–170.

Linklaters (2014). Capacity mechanisms. reigniting europe's energy markets. Technical report.

Perez-Arriaga, I. J., editor (2013). *Regulation of the Power Sector*. Springer, 1st edition edition.

Perez-Arriaga, I. J. and Meseguer, C. (1997). Wholesale marginal prices in competitve generation markets. *IEEE Transactions on Power Systems*, 12(2).

Rodilla, P. and Batlle, C. (2012). Security of electricity supply at the generation level: Problem analysis. *Energy Policy*, 40:177–185.

Schweppe, F. C., Caramanis, M. C., Tabors, R. D., and Bohn, R. E. (1987). *Spot Pricing of Electricity*. Kluwer Academic Publishers, Boston.

Sioshansi, F. P. (2008). *Competitive Electricity Markets. Design, Implementation, Performance*. Elsevier, CA.

Vazquez, Carlos, M. R. and Perez-Arriaga, I. J. (2002). A market approach to long-term security of supply. *IEEE Transactions on Power Systems*, 17(2).

Exhibits

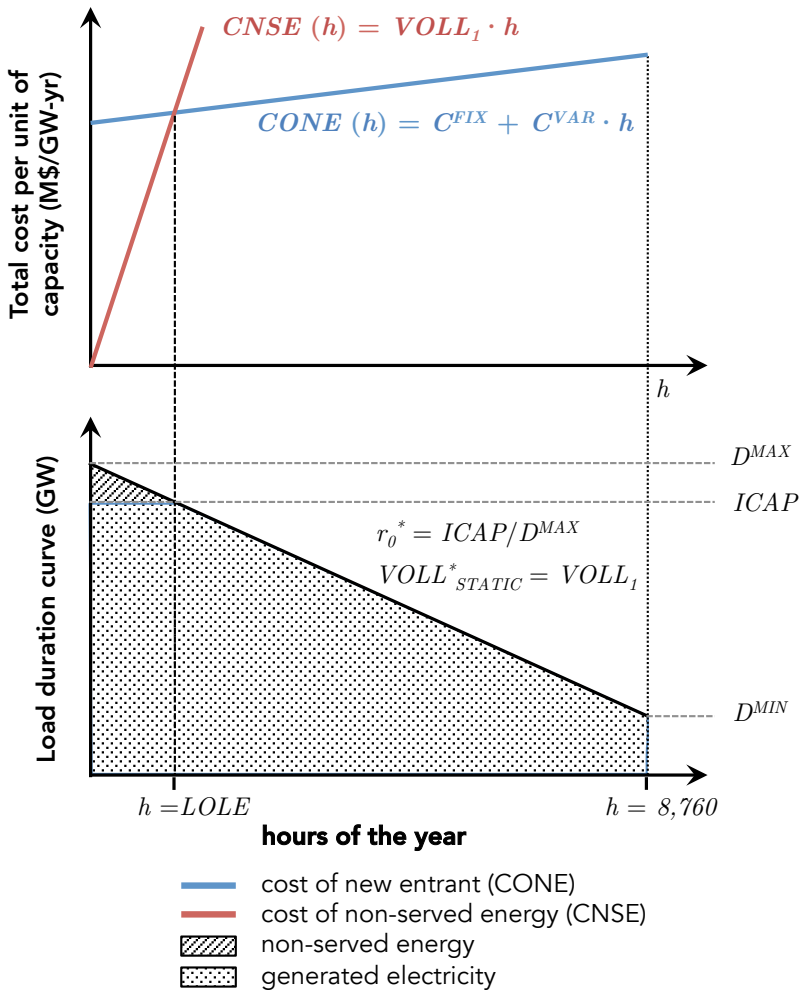

Exhibit A: Illustration of the single-technology generation expansion model, and the process to determine the optimal value-of-lost-load ($VOLL$) that yields the desired reliability level, given by a targeted number of hours with non-served energy. $CNSE$ denotes the cost of non-served energy; $CONE$ denotes de cost of new entrant, equivalent in this model to the cost of generation capacity; h denotes hours; C^{FIX} denotes the annualized fixed costs of generation; C^{VAR} denotes the variable costs of generation; D^{MAX} denotes the peak demand; D^{MIN} denotes minimum demand; $ICAP$ denotes installed capacity; $LOLE$ denotes the number of lost load events; r_0^* denotes the optimal capacity demand ratio at the time of investment yielding the required $LOLE$; and $VOLL^*_{STATIC}$ denotes de optimal value of lost load in the static analysis.

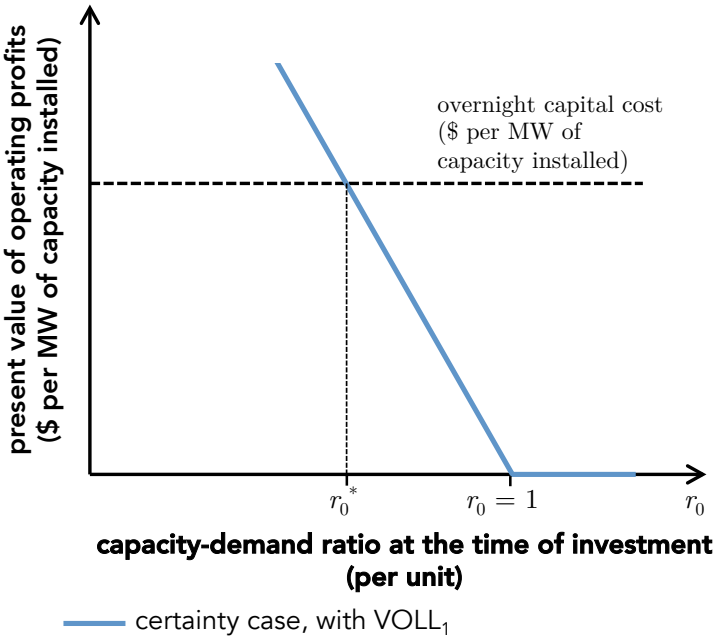

Exhibit B: Illustration of the present value of operating profits earned by each unit of installed generation capacity as a function of the capacity-demand ratio at the time of investment, and the optimal capacity-demand ratio, r_0^*.

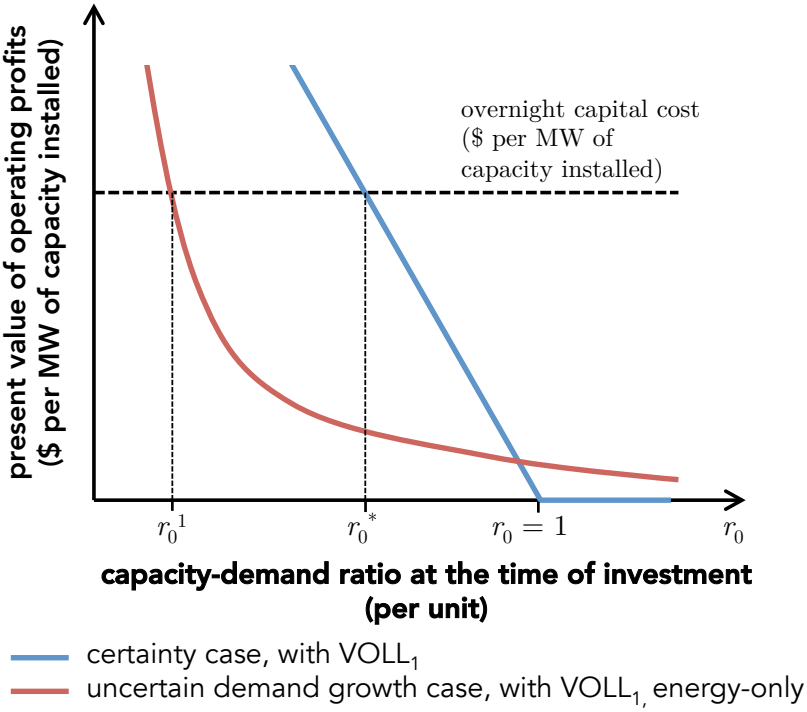

Exhibit C: Illustration of the present value of operating profits earned by each unit of installed generation capacity as a function of the capacity-demand ratio at the time of investment, with uncertainty on demand growth. r_0^1 denotes the new capacity-demand ratio equilibrium, and $r_0^* - r_0^1$ represent the discrepancy between the optimal capacity-demand ratio and the case with uncertainty in demand growth.

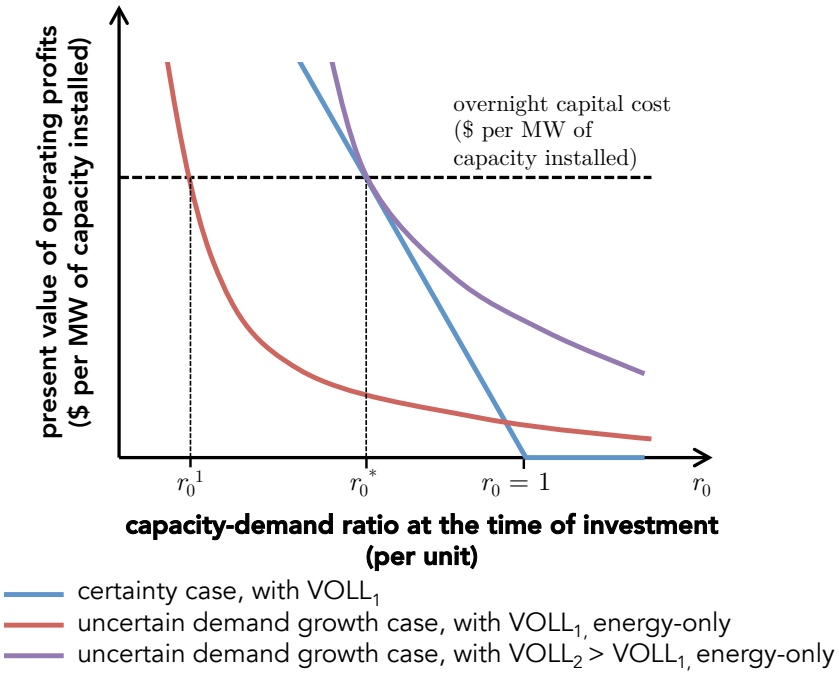

Exhibit D: Illustration of the effect of the introduction of a larger $VOLL$ ($VOLL_2 > VOLL_1$) on the present value of operating profits, in an energy-only market. The increase of $VOLL$ leads to the optimal level of investment under demand growth uncertainty.

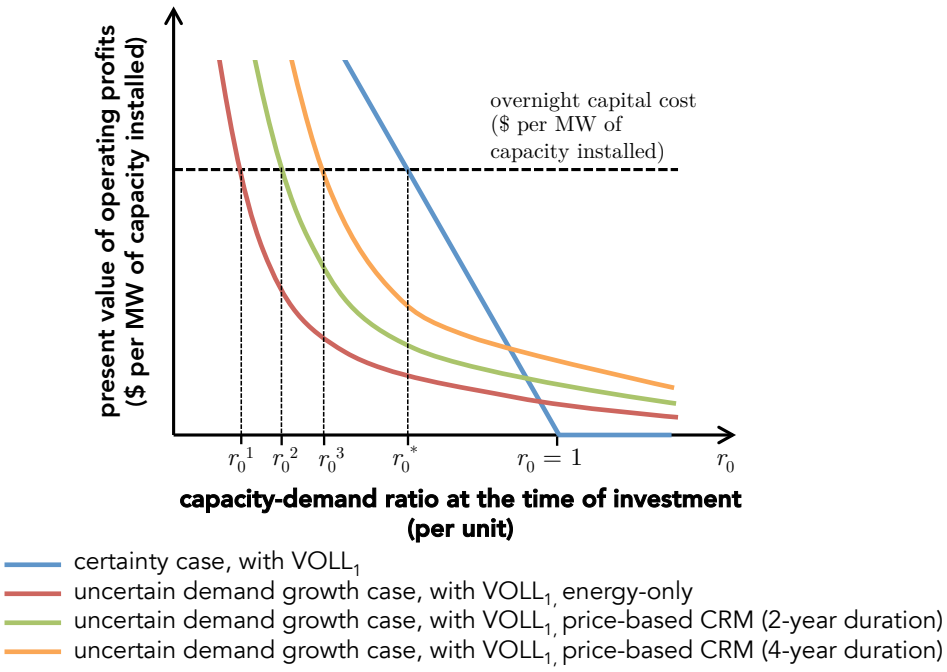

Exhibit E: Illustration of the effect of the introduction of a price-based CRM with different contract durations that lead towards the optimal level of investment. Larger contract durations can potentially shift the capacity-demand ratio in equilibrium to the optimal ratio.

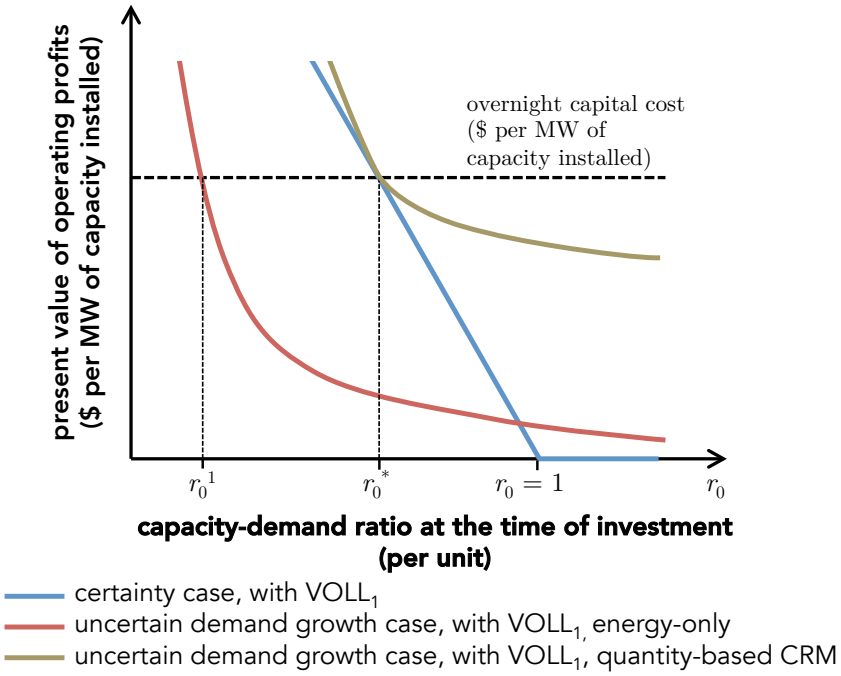

Exhibit F: Illustration of the effect of the introduction of a quantity-based CRM where investors can internalize uncertainty in their bids on the present value of operating profits, producing a new equilibrium that reaches the optimal capacity-demand ratio.