Comparing Capacity Market and Payment Designs for Ensuring Supply Adequacy

Robert Stoddard* 
CRA International
RSStoddard@crai.com

Seabron Adamson*
CRA International
SAdamson@crai.com

Abstract

Capacity payments and capacity markets are one mechanism for ensuring that sufficient marginal generation is built to meet peak demands with an adequate reserve margin. The European Union experience with such mechanisms has focused on capacity payment mechanisms, while in the United States the trend has been towards quantity-based mechanisms, especially in the form of locationally-specific, centralized procurement auctions on a forward basis, with the grid operator securing capacity commitments three or four years ahead and allocating costs based on realized peak load. Historically, many of the European and U.S. capacity market designs have been generally ineffectual. Recent American designs have evolved towards contractual mechanisms that guarantee adequate generation investment but at potentially higher costs.

1. Introduction

The need for generators to receive revenues beyond clearing energy prices has been recognized in the development of power markets from the beginning. The spot pricing formulation of Schweppe et. al. included a generation quality of supply term as part of the spot price, which was effectively a price to ration demand when generating capacity became scarce [1]. Under this short-run marginal cost (SRMC) pricing framework, this component of the spot price would also be paid to generators, helping recover the total costs of generation (including the fixed and sunk costs of building generating units).

Constraints on market prices – such as price caps and other controls on market power that may limit scarcity price signals - are often quoted as a reason why a separate capacity mechanism is need to support resource adequacy and security of supply. In many countries, including the United States, the case for specific capacity mechanisms has also been driven by policymakers’ need for assurance that a market-based system of generation investment can guarantee the reserve margin needed to protect reliability.

2. Need for capacity mechanisms in power markets

In an energy-only market there may be insufficient revenues (in expectation) for investors to build sufficient peaking plant to meet a prudent reserve level. AS shown in Fig. 1, for a specific cost of new entry investors will build plants to an “economic” reserve margin level, producing energy market revenues per kilowatt of $R^*$. However, this reserve level will typically be below the engineering reserve margins considered necessary to protect reliability to national and international standards, which produces net energy market revenues of $R^\wedge$. The difference $R^* - R^\wedge$ is often referred to as the “missing money” – the additional revenues needed to support a reliable system at a fixed reserve margin.

---

*This work solely represents the opinions of the authors and may not represent the views of CRA International or its clients.
capacity is unknown in advance. In quantity-based mechanisms, a fixed amount of capacity is pre-determined, based on expected load and a required reserve requirement. Under these mechanisms, however, the price is unknown. There are strong analogues between choosing reliability mechanisms and the price-versus-quantity debates of environmental economics [2].

3. European capacity payments

The quality of supply framework – which in practice reflected a capacity payment that became significant in periods of tight supply – was reflected directly in the design of the first European power market in England and Wales. Under the Pool rules, the capacity component of the Pool Purchase Price was equal to:

$$\text{LOLP} \times (\text{VOLL} - \text{SMP})$$

where LOLP was the calculated Loss of Load Probability, VOLL was an administratively determined value of lost load, and SMP was the system marginal price in the half-hour periods [3]. The LOLP capacity payment, as noted by several authors, was widely viewed as too easily affected by the bidding decisions of major generators [4]. The LOLP payment, along with the rest of the Pool rules, was scrapped with the introduction of the New Electricity Trading Arrangements in England and Wales.

3.1. The Spanish capacity payment mechanism

The Spanish market has included a capacity payment from the start. The garantía de potencia was established by Real Decreto 2019/1997 in 1997. Under this structure, suppliers and distributors pay charges that form a “fund” established in advance from which capacity payments are paid. These funds are paid out to all generators that sell their energy in the market in proportion to the installed capacity. The rules were later changed to exclude nuclear plants.

This structure had a number of fundamental weaknesses [5]. First and foremost, the capacity payments were unrelated to supply and demand for capacity, and hence could not provide economic signals for investment. Second, there was no real defined capacity product, so generators were paid regardless of their actual contribution to system reliability. Finally, there was no defined capacity margin and hence no real guarantee that reserves would be provided. The mechanism has recently been modified.

3.2. Capacity payments in the All-Island market

The All-Island Market for Electricity (AIME) encompasses the Republic of Ireland and Northern Ireland. Under the Capacity Payment Mechanism rules, an annual “capacity sum” is calculated from the product of the Capacity Requirement (load plus reserves) and the cost of the Best New Entrant (BNE) generator [6].

Under the CPM rules, this amount is split into monthly quantities, based on allocation factors. This monthly sum is then divided into three “pots”. Each of these “pots” is allocated across generators using different metrics: annual load forecast, the ex-ante LOLP, and the ex-post LOLP.

AIME began operations very recently, so there is little practical experience of its effectiveness. Like other such designs, the CPM appears to be best suited for period in which capacity is tight, as aggregate revenues appear always to be set by the cost of new entry. Determining the costs of the BNE may also prove contentious, as have similar U.S. exercises.

4. Installed capacity markets in the U.S.

The three power markets in the Northeastern United States – PJM, the New York Independent System Operator (NYISO) and ISO New England – adopted a different construct for ensuring resource adequacy. Each of these markets had antecedents in benefits-sharing “tight pools” that predated restructuring. These pools commonly had fixed reserve margin requirements and “capacity ticket” mechanisms that were used to allocate responsibilities for meeting the margin to the participating load-serving entities (LSEs).

4.1 Installed capacity markets

Each “capacity ticket” was for one unit of installed capacity (ICAP). The operation of the ICAP market was in theory simple. Each owner of qualifying generation could sell ICAP up to its total installed capacity, either bilaterally or in ISO-run auctions. Each
LSE needed to acquire ICAP up to its forecast demand plus the set reserve margin. So, for example, if the set reserve margin was 15%, then an LSE supplying 100 MW of load would need to acquire ICAP from generators equal to 115 MW or face a penalty. The ICAP market was independent of the energy market and selling ICAP did not require the generator to provide energy at any time. In concept, it was believed that a price for installed capacity would emerge, plus an additional revenue stream to support resource adequacy.

While this simple trading mechanism worked acceptably when substantially all capacity was owned by regulated utilities, market liberalization and the subsequent divestitures of utility-owned generation revealed a number of flaws [6]. Market power has been viewed as an issue, given the vertical demand curve originally used in the capacity markets. The product definition of ICAP was also problematic, and gave poor incentives for plants receiving capacity payments to be available when needed.

One early improvement to the market design was to shift the product from installed capacity, ICAP, to unforced capacity, UCAP. The ICAP value of a typical resource was simply equal to its tested seasonal rating. This metric does not include any adjustment for the reliability of that resource in actual operation, however, and therefore provided no direct financial incentive to maintain or improve a unit’s reliability when called. PJM developed a metric, the Effective Forced Outage Rate under Demand (“EFORD”), based on historical forced outages of a resource to approximate its likelihood of being on forced outage during periods of high system loads. The UCAP value of a resource is calculated as the product of its ICAP value and \(1 - \text{EFORD} \). This shift in the capacity product definition had the clear effect of decreasing forced outage rates and, therefore, increasing system reliability with a fixed set of assets.

### 4.2. Variable resource requirements

A solution to the bipolar pricing mechanism was proposed in the form of a variable resource requirement (VRR), often known colloquially as the “demand curve” for capacity. In simple terms, the VRR construct changes the price of capacity to be bought by the ISO in a fixed and pre-determined relationship to the quantity offered.

This relationship is illustrated schematically in Fig. 3. At the target reserve margin \(Q^*\) the price is supposed to match the cost of new entry (CONE) for a peaker, after subtracting net energy and ancillary services revenues. As the quantity offered increases, the marginal benefit of additional reserves is less and the price falls, until at a reserve margin well above the target the price of installed capacity falls to zero. At reserve margins below the target, the marginal benefit is higher and so the ICAP prices rises. Prices were effectively capped at fixed levels (the flat part of the VRR curve in Fig. 3) in the implementation of this concept in the NYISO.1

The implementation of variable resource requirements helped remove some aspects of the “bipolar” pricing problem which plagued the U.S. ICAP designs, by adding a measure of (administratively determined) demand elasticity into the capacity equation. Supply elasticity, however, remained low as new entrants could not enter the

---

1 In the NYISO design, the installed capacity market design included a locational aspect (LICAP), with New York City, Long Island and Rest of State zones. Each LICAP zone had its own “demand curve” and parameters are defined which determine how much capacity can be imported into the constrained New York City and Long Island zones.
market over the (monthly) horizons) used in the ICAP markets.

Although this VRR-based market design is a marked improvement over the ICAP design, it is not without its own problems. The two most serious issues are with setting the VRR parameters and with the exercise of market power, both by buyers and sellers.

![Diagram](image)

**Fig. 3. – Variable resource requirement for capacity**

Market-clearing prices in a VRR market depend critically on the particular VRR function. Although various functional forms have been debated [7], dictates of simplicity have led all VRR proposals to be piece-wise linear, making three parameters of particular relevance: \( \text{CONE} \), the quantity at which the capacity price falls to zero (the x-intercept, or the “zero-crossing point”), and the price cap. The price cap and zero-crossing point are only loosely bound to any economic or reliability principles. The price cap must be high enough to ensure that the actual value of \( \text{CONE} \), as determined in the marketplace, is below the price cap, as otherwise the price signal from the capacity market will be too low to attract new entry. The zero-crossing point determines the elasticity of the VRR curve and, consequently, has significant effects on the competitive dynamics of the market design. The closer the zero-crossing point is to the minimum resource requirement point, the more “bi-polar” the market will become. Increasing the zero-crossing point, however, has the effect of increasing capacity payments by LSEs during periods of surplus capacity, and although it has been shown that the long-run cost of a more elastic VRR curve is lower than with a more inelastic one [8], there has been tremendous political pressure in designing these VRR curves to reduce the zero-crossing point.

The key parameter of a VRR curve is the administrative estimate of \( \text{CONE} \). This estimate must try to replicate the financial calculations of a merchant generation developer and, therefore, assesses three elements: construction costs, financing costs, and expected future net cash flow. Of these, construction costs is arguably the least controversial to measure, and the three U.S. power markets all undertook this step by retaining an expert consultant to price out the engineering costs, benchmarking the results to recent observable installation costs. Financing costs are mechanically straightforward to compute, but doing so requires agreement on cost of equity, debt financing structure and interest rates, residual value, and method of levelization. Expected future net cash flows are a highly contentious element of the estimate. Peakers are used as the benchmark to reduce the importance of this element in the estimate, but consequently the highly volatile nature of the earnings stream is difficult to model or predict. PJM simply relies on historical earnings of similar units, with no attempt to adjust for changes in market conditions. NYISO’s consultant has used a more sophisticated probabilistic method. The chief difficulty with the process, however, is that there are clear (short-term) winners and losers: loads pay less if \( \text{CONE} \) is lower, while generators seek a higher \( \text{CONE} \) to increase their payment. Depending upon the governance structure of the market, it may be politically very difficult to agree on a level of \( \text{CONE} \).²

The second major issue with a VRR curve is with market power. Although the VRR curve introduces some elasticity into the “demand curve,” all of the VRR curves adopted in the U.S. are still highly inelastic, i.e. total payments to generators decline (sharply) as the quantity in the market increases. Consequently, there may be strong incentives for large sellers and large LSEs to engage in the exercise of market power. A single large seller can profitably withhold capacity, increasing both the overall capacity payments and its own payments. The potential for such economic withholding has long been discussed in New York City, where three suppliers control most of

---

² Even in the face of unrefuted evidence that construction costs of new power plants had risen materially between 2004 and 2008, two key stakeholder committees of PJM voted down an increase in the \( \text{CONE} \) value. The Board of Managers decided to file the increases with FERC nonetheless, on recommendation from PJM senior staff that such increases were essential to the sound operation of the market. The matter is being contested at FERC.
the capacity. Although the three are subject to a must-offer requirement and an offer cap, it has been alleged that one supplier has consistently offered most of its capacity at that offer cap price even when the knowable outcome is that not all of that capacity would clear, effectively supporting the clearing price at the offer cap. Although this action was consistent with the NYISO market rules, it has led to a reexamination of whether the NYISO design remains appropriate, at least for load pockets like New York City. Conversely, the steep VRR curve also creates incentives for a large LSE to contract with new supply to enter the market in advance of need. By sponsoring new build directly into the market, the LSE must pay the full contract price to new resources but can suppress the market-clearing price it must pay for existing resources. Suppliers have alleged that the two regulated New York City LSEs have engaged in precisely this behavior when they underwrote 1,000 MW of new capacity construction. Although market rules mitigating these exercises of market power by both sellers and buyers can be adopted, economists and market designers continued to search for an alternative that would bring more market forces to bear.

4.3. Forward contracting for capacity

The solution adopted in some U.S. markets has been to shift towards forward acquisition of capacity, over a period long enough so that new entrants (that is, developers with firm plans for new plant but without currently operating generation) can bid in supplies and then build the units if they are successful in the capacity procurement auction. This has the effect of flattening the supply curve, as shown in Fig. 4.

Two U.S. ISOs have recently restructured their capacity markets to operate on the basis of auctions for forward capacity. The Reliability Pricing Model (RPM) in PJM and the Forward Capacity Market in ISO New England have both begun operations in the last year, after several years of development. Both designs have features that are influencing current capacity market developments elsewhere in the U.S.

![Fig. 4. - Contracting for capacity in advance flattens the supply curve](image)

5. The RPM and FCM designs

By moving the time when the primary market is cleared from a few days before the delivery period to a three or four years prior, the RPM and FCM designs are intended to improve market outcomes by allowing direct competition between incumbent suppliers and potential suppliers. The potential benefits are substantial. Structural market power of incumbents is diluted as the market becomes contestable [9]. Likewise, market power of buyers can be muted by retirement of less-efficient generation in response to new construction underwritten by LSEs. New resources can displace economically inefficient, older resources in an orderly way, minimizing the potential reliability risks often associated with unit retirement.4 Transmission planning can be better meshed with generation development and retirement, including the direct economic evaluation of substitution between transmission and active generation in load pockets.

The first capacity market design with a multi-year forward commitment was developed for a Central Resource Adequacy Market for PJM, NYISO, and ISO New England [10]. This attempt to coordinate the capacity markets across the entire Northeast U.S. region was aborted, however, leaving each market operator to develop its own markets.

---

3 New York City has a peak demand of about 13,000 MW, so it is large by any conventional standard. Only about 5,000 MW can be imported through the transmission system, however, creating localized market power for the large suppliers in the city.

4 Merchant generation in the U.S. generally has the right to retire with fairly short notice period, typically a few months. When a retiring unit is needed to maintain system reliability, however, it may negotiate a Reliability Must Run ("RMR") agreement with the market operator. An explicit goal of modern capacity markets is to eliminate nearly all RMR contracts, because the market operator will have had several years notice of a unit’s retirement and can plan accordingly.
5.1. PJM Reliability Pricing Model Design

PJM was the first U.S. market to begin development of such a multi-year resource adequacy design. The resulting RPM design evolved the PJM design in four critical ways.

First, the RPM capacity product was defined locationally. Previously, all capacity in PJM was deemed to be deliverable anywhere to the extent it was deliverable to the transmission grid. In a market with over 170,000 MW of generation spanning over 1,120 km, this premise was, unsurprisingly, false. A locational attribute was needed to provide the correct price signals to build where capacity was in short supply, rather than where construction costs were lowest.

Second, although the capacity product definition remains UCAP, payments are linked to a resource’s availability during the approximately 500 hours with the highest expected loads. The intent is to provide an even greater incentive for resources to be available during peak hours, when energy price caps mute the market signal for availability.

Third, it incorporates rules to mitigate the exercise of potential market power for both suppliers and loads. Suppliers of incumbent resources are limited to bidding their demonstrated going-forward costs for nearly all capacity. New resources offered below cost because of contract support from a load-serving entity have a limited ability to set the market clearing price. Although some have argued that the mitigation on supply is too tight and on load, too lax, the clearly defined mitigation has the benefit of reducing market uncertainty.

The most important new feature of RPM, however, is the market-clearing mechanism. PJM runs a Base Residual Auction (“BRA”) approximately three years prior to each planning year (which runs from June to May). This auction clears all capacity in the market, including self-supplied resources, using a VRR.5

Including both forward procurement and a VRR introduces a certain degree of redundancy in the market design. The VRR mechanism was introduced primarily to stabilize the price formation in the market created by having highly inelastic supply. In a forward market, however, supply is no longer inelastic; entry and exit can occur before the delivery year, based on the outcomes of the auction. Including the VRR also brings with it all of the economic and political challenges of developing and agreeing on the VRR parameters, not merely at the start of market but also as costs and other market conditions evolve.

5.2. New England Forward Capacity Market Design

After the collapse of the three-market design initiative, New England set out to improve its capacity markets. It initially chose a market structure essentially identical to that of the NYISO, though with improved market power mitigation. The regulators in the six New England states, however, demanded a different direction, consistent with the earlier three-market initiative, and so the Forward Capacity Market (“FCM”) was negotiated.6

Like the RPM design, FCM is a market for capacity run about three years in advance of the planning year, and it contains appropriate market power mitigation on both suppliers and loads. There are three key differences in the design, however.

First, the FCM has no VRR. The quantity cleared in the market is not more than the required reserve margin (although it may be less under extraordinary conditions). The market relies, therefore, primarily on the elasticity of competitive supply to produce reasonable prices.7 Discarding the VRR curve allows the market price to reflect the actual competitive cost of capacity in the market at all times, even if the true CONE differs substantially from an administrative estimate of CONE.

Second, the capacity product sold is a hybrid physical and financial commitment. Like all other U.S. capacity markets, a resource cleared in the FCM is committing to offer that resource into the day-ahead energy market whenever it is physically available. While this obligation is transferable, it must be transferred to another qualified capacity resource available for dispatch by the market operator. In this sense, the capacity obligation is physical. A capacity

---

5 Entities have an option to recuse themselves from the RPM if they take full resource planning responsibility for all load in their service area for five-year periods. This accommodation was reached to allow regulated LSEs to operate within their current regulated structure.

6 One of us (Stoddard) represented the Capacity Supplier coalition in this settlement proceeding.

7 There are several other market rules that contribute to price formation stability, including rules to address market power and “lumpy” investment in small load pockets. There is also a temporary price cap and floor to ensure reasonable outcomes during market startup.
resource is also selling a financial call option. This option has a strike price equal to the dispatch cost of a marginal resource (deemed to have a heat rate of 22,000 Btu/kWh), and it is struck against real-time prices. Capacity obligation holders are responsible for payments under this option regardless of whether their resource was available for dispatch, making it very similar to the reliability options described in the literature [11] [12].

Third, FCM is built on the ICAP product definition, rather than the UCAP product used elsewhere. ISO New England did not believe that the EFOR metric provided meaningful incentives to be available when a resource was most needed. The FCM design, therefore, contains no penalties for unit outages during normal system operations, but it includes very stiff penalties if a unit is unavailable when the system is short of operating reserves, ranging up to 10 percent of the annual payment per day and 21 percent per month. These penalty charges are in addition to the call option payments that will necessarily occur in these shortage periods.8

Collectively, these differences between the FCM design and the RPM bring the FCM closer to a market-like mechanism for assuring resource adequacy. Administrative estimates of CONE and other VRR parameters have little influence in the market outcomes. While the quantity of capacity procured is determined administratively in either market, based on reliability needs assessments, the FCM design does not, under normal market outcomes, make a tradeoff above or below that mark based on costs. The sharper “pay for performance” of the energy call option and targeted availability metrics is a closer approximation of the timing of scarcity rents that are missing from the energy market, although it remains to be seen whether the increased financial risk to (and consequently, offer prices from) generators will be generally acceptable.

5.3. Outcomes of RPM and FCM to Date

PJM and ISO New England have both implemented their respective RPM and FCM markets.

Because the PJM RPM design includes a VRR curve that stabilizes prices even when new entry cannot, PJM was able to run its market even with much less than the full three-year commitment period. Since the market tariff was approved in December 2006, PJM has run the BRA for four delivery years: 2007/08 through 2010/11. For the first three of these years, the market outcomes have been in line with market fundamentals, with clearing prices higher in the constrained eastern markets than in the relatively resource-rich western regions.

In 2010/11, however, prices in the east declined and converged with those in the west. These moves appear to be primarily the result of expected future transmission upgrades coupled with a decline in the administrative CONE resulting from increased expected energy margins. This CONE value understates capacity construction costs, however, so prices in the 2011/2012 BRA are expected to increase significantly.

RPM has met with reasonable success in attracting new resources where needed. Its success has been clearest in attracting new demand-side resources, which are those customers able to curtail usage in response to operator instructions: 1,373 MW of new demand resources have cleared over these four auctions [13]. PJM also reports that 3,228 MW of capacity resources have withdrawn their request to deactivate, postponed retirement, or been reactivated since the RPM design was agreed. Although there have been relatively fewer new generating units added, that behavior is consistent with market prices below the level new resources would require. Nor is it an inefficient outcome, if capacity from demand resources or existing resources is both lower in cost and sufficient in quantity.

The first Forward Capacity Auction (“FCA”) of the New England FCM was held in early February for the 2010/11 planning year. It was highly successful by many measures. At the time the FCM was agreed, there was less than 1,000 MW of generation in the interconnection queue, much of which was unlikely to be built for various siting and technical reasons. Following the adoption of the FCM, more than 13,000 MW was added to the queue, and consequently bidding into the FCA was very robust. ISO New England received offers from 39,155 MW of resources, competing to meet the 32,305 MW capacity requirement. The auction cleared more than 1,813 MW of new supply-and-demand-side resources, adding 1,188 MW of demand-side resources to the existing base of 1,366 MW. Collectively, these results imply that fully 10% of New England’s peak load will

---

8 The New England ancillary services market design includes administrative “constraint violation penalties” that add $850/MWh to $1000/MWh to the real-time prices of energy and reserves when there is a reserve shortage.
available for dispatch to market operators, the highest amount of committed load reduction on any U.S. system and, quite possibly, in any organized market.

6. Concluding Observations

Both of the forward capacity markets now implemented in the United States have been remarkably successful in their primary objective: ensuring resource adequacy, both in the pool overall and in transmission-constrained regions. New capacity resources of all types have cleared in these markets, and the interconnection queues now include many proposals for all types of resources. Of particular note is the high degree of demand-side participation in these auctions, demonstrating clearly that, if customers are given a clear economic signal of the cost of their electricity consumption and the ability to act on that signal, the demand for resources is in fact flexible.

Both designs, however, have attracted many complaints about particular details of their implementation. Buyers in the RPM market complain that the sloped VRR curve forces customers to buy more capacity than is required in some years, notwithstanding theoretical arguments that doing so saves costs in the long run. Both suppliers and buyers in PJM complain about the particular parameters of the VRR curve, predictably asserting that the curve is set too low or too higher, respectively. The New England FCM has, overall, raised fewer concerns, a predictable outcome of the relative lack of administrative parameters in the design. There have been issues, however, with the stability of the price formation when many load-serving entities are self-supplying new capacity in the market at zero-price bids. Overall, however, this FCM design appears to the authors to be more attractive as a framework for future capacity market developments, as it is more consistent both economic theory and political reality.

7. References


8. Biographies

Robert B. Stoddard (b. 1961) received his undergraduate degree in economics from Amherst College (1983) and his graduate economics degrees from Yale University (1986, 1990).

Mr. Stoddard is a Vice President in the Energy & Environment practice of CRA International, where he leads the Regulation & Litigation group. He played a significant role in the design of the New England and PJM capacity markets and represents both U.S. and European clients in other matters concerning resource adequacy and long-run investment in power markets.

Seabron C. Adamson (b. 1964) received the undergraduate and Master degrees in Physics and Applied Physics (1986 and 1987) from the Georgia Institute of Technology. He also holds the M.S. degree in Technology and Policy from the Massachusetts Institute of Technology (1992) the M.A. degree in Economics from Boston University (2007).

Presently he is Vice President in the Energy & Environment practice of CRA International, where he leads the Enterprise and Asset Investment group. He has advised clients on resource adequacy and capacity markets in the United States and the European Union.