Locational Electricity Capacity Markets: Alternatives to Restore the Missing Signals

In the absence of a properly functioning electricity demand side, well-designed capacity payment mechanisms hold more promise for signaling the value of capacity than non-CPM alternatives. Locational CPMs that rely on market-based principles, such as forward capacity auctions, are superior to cost-based payments directed to specific must-run generators, as CPMs at least provide a meaningful price signal about the economic value of resources to potential investors.

Amparo D. Nieto and Hamish Fraser

Amparo D. Nieto *is a Senior Consultant at NERA Economic Consulting in Los Angeles. She specializes in the economic regulation of the electricity industries and has provided extensive advice in the areas of dynamic retail pricing and marginal cost modeling, design of energy and capacity markets, and transmission pricing. She holds an M.A. in Economic Analysis and Public Finance from the Fiscal Studies Institute in Madrid, and a B.A. in Economics from Carlos III University in Madrid.*

Hamish Fraser *is a Special Consultant at NERA, based in its New York office, who specializes in the detailed design and analysis of electricity markets.*

The authors would like to thank their colleagues Gene Meehan and Jonathan Falk for their valuable comments.

I. Introduction

The reliability problems in an interconnected electricity system ought to be solved or avoided through an efficient combination of investments in transmission, generation, and/or demand-side alternatives. However, if market participants are to make efficient investments and usage decisions, the market must send the proper price signals. Prices need to reflect the true economic value of incremental energy and capacity in any *time period* and *location*. Unfortunately, in most electricity markets, this is not the case.

Even in restructured markets, regulatory arrangements to ensure capacity adequacy are generally still needed, at least in the near term, and the theoretical and practical arguments for establishing electricity capacity payment mechanisms (CPM) have been widely accepted in the U.S., Europe, and elsewhere. Most

 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

markets are incomplete because they are missing important players: the end users. Demand response has been slow in coming so only a small volume of demand is exposed to market prices or is able to signal how much users are willing to pay for a given level of reliability. Most customers do not ration their demand when market prices are high simply because they do not see these prices – their rates reflect generation costs averaged over long daily or seasonal periods. This averaging of prices is due not only to equipment constraints (lack of interval metering or real-time communication) but also to the perception that customers will not be willing to accept even limited levels of exposure to actual short-term market price volatility.[1]

The lack of a meaningful demand side leads to reliability concerns and market price volatility. When economic signals of scarcity conditions are not delivered to end users, the demand curve is inelastic, and with nothing to constrain prices, market energy prices can spike to very high levels. Energy regulators typically react by imposing price (or bid) caps in the energy market[2] so that in hours when the system is most stressed, there is no scarcity price signal to elicit efficient generation entry or demand-response initiatives. In addition, low volatility risk in the energy market severely reduces the incentive for retailers to sign long-term bilateral contracts with generators.[3] In absence of new entry, the system reserve margin will continue to decrease while the associated shortage costs increase.

Until demand response is facilitated and market prices reflect the underlying marginal costs of energy and capacity, a ‘‘second best’’ solution (i.e., a CPM) is needed to mitigate this social deadweight loss. The next question for economists is: What exactly is this second best? There is no universal agreement on the ideal CPM design for electricity markets, and different schemes work better in different conditions. However, one problem common to most of the CPMs implemented to date worldwide is that there is generally a single capacity price for an entire region or country. Such CPMs fail to acknowledge the *locational* value of capacity. This lack of a locational dimension in the capacity mechanism may worsen the capacity and transmission problems in specific locales, even while it resolves the problems in the aggregate.

The lack of a meaningful demand side leads to reliability concerns and market price volatility.

The theoretical arguments for locational marginal prices (LMPs) for energy are well established and widely implemented. Location-specific energy pricing provides nodal information on energy losses and congestion. It is justified as a market-based means of promoting short-term allocative and productive efficiency, as well as long-term efficiency in the location of generation, transmission, and demand. LMPs signal generators where their power will be most valuable, merchant transmission investors where new lines will be needed, and industrial firms where power consumption is less costly. A locational capacity payment mechanism, therefore, should not be particularly controversial – at least not conceptually.

In theory, the economic value of capacity as part of a reliability solution is defined by the marginal ability of that capacity to reduce loss of load expectation (LOLE). LOLE analysis looks at the volatility of demand and reliability of generator capacity to determine how many hours a year, on average, demand is likely to exceed capacity. Assuming the output of a hypothetical plant in the region is deliverable on an equivalent basis when and where it is needed, regardless of where the plant is located, then the value of capacity will be constant across a region. In practice however, some plants may be unable to operate at full output at peak times due to network constraints. With a single capacity price set for the entire region, generation entry in load pockets or import-constrained areas does not occur

 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

and the transmission constraints worsen over time, as load grows. These areas become increasingly vulnerable to high peak-demand periods due to extreme weather conditions or unplanned outages of generation and transmission capacity. The U.S. illustrates these problems well. In the first years of restructuring, plant construction boomed and for the most part aggregate capacity levels were well above the levels required. But the adequacy of generation capacity has not been uniform within each region, as most plants were built in low-cost areas rather than in load pockets where they are most needed.

In the rest of this article we explore the range of alternatives that have been implemented both in the U.S. and abroad to provide incentives for solving locational problems. First, we examine the more traditional methods outside of a centralized CPM construct. We then look at types of centralized CPM schemes currently considered in a number of U.S. markets as part of their regional resource adequacy plans. We evaluate these methods from the point of view of the incentives provided to market participants, and their potential impact on the efficiency in the market.[4]

II. Locational Capacity Solutions Outside a CPM Construct

Electricity sectors around the world have implemented methods to provide locational incentives long before the introduction of any locational CPM constructs. Methods include reliability must-run (RMR) contracts and special bidding arrangements, local ancillary services markets, incentives for merchant transmission, locational access charges for generators, and non-wires solutions. None of these methods link remuneration of the generation, transmission, or demand response resources to consumers' valuation of lost load. As a result, the approaches do not reflect that consumers would reduce demand if they saw the higher prices associated with capacity shortages, and ultimately are not likely to be least-cost solutions.

The main attraction of the RMR mechanism is that contracts are only signed with those generating units that need them.

A. Reliability must-run contracts

In instances where capped energy market revenues fail to compensate the operating and avoidable fixed costs of necessary local resources, typically old high-cost peaking units with low capacity factors, cost-based payment mechanisms can be applied on a case-by-case basis to keep these units available.[5] In U.S. markets such as New England, PJM, and California, typical compensatory local reliability mechanisms have taken the form of reliability must-run contracts between the ISO and selected units located in load pockets. The main attraction of the RMR mechanism is that RMR contracts are only signed with those generating units that need them and where those units are also needed for system reliability purposes.

Nevertheless, there are a number of problems with RMR contracts. In particular it may be difficult to identify those units which: (a) have the need for a particular level of price guarantee in order to remain in operation; and (b) must be kept in operation in order to maintain acceptable levels of reliability. A central entity should make a very detailed and potentially extraordinarily complex unit-by-unit analysis of the entire system, including the avoidable fixed- and variable-cost levels of individual units, the reliability consequences of having/not having each unit, the strategic response by market participants to alternative compensation or closing for each unit, the market power implications of each options, etc. In practice, the selection of RMR units is usually more pragmatic and *ad hoc*, and this can lead to disputes.

Further, the mixture of reliance on cost-based regulation for some generators and market-based forces for others can interfere with the dynamic efficiency of the generation markets. By providing the RMR generators with out-of-market contract payments, as opposed to letting locational market prices signal the higher value of capacity in a constrained area, market forces are unable to properly fulfill their function of signaling the true scarcity value of energy to either load or generation. Generally, neither the load in the constrained area nor potential entrants are aware of the extra cost associated with RMR contracts, since these costs are typically spread to all consumers in the region in the form of uplift charges.[6] Over time, the uneconomically low market prices within load pockets may over-stimulate demand, requiring ever higher out-of-market generation and transmission expenditures.[7]

B. Special bidding provisions (''PUSH'')

In 2003, PPL and Devon Power applied to FERC for RMR status on some specific units in New England. ISO-New England required these units to remain available for reliability purposes even though they were not viable at prevailing market prices. As part of the prevailing complex rules relating to market power mitigation in designated congestion areas, price caps had been imposed based on the annual cost of a new combustion turbine unit divided by the number of hours it was expected to operate during the year. Recognizing the need for cost recovery, FERC changed the rule and introduced a new mechanism called ''peaking unit safe harbor'' (PUSH) as a temporary measure in New England, pending an alternate and more permanent solution.

PUSH was designed to allow a generator that had operated at a capacity factor of 10 percent or less in 2002 a safe harbor bid price based on the sum of its variable cost and fixed costs. The fixed cost component for 2003 was calculated by dividing the unit's annual fixed costs by the number of MWh it generated in 2002. The generator could therefore recover its costs in 2003 if it ran for the same number of hours (and had the same costs) as in 2002. These prices were also allowed to set the LMP.

In theory, the PUSH approach is more efficient than RMR contracts. However, it has encountered problems of its own.

In theory, the PUSH approach is more efficient than RMR contracts in that it allows the market price in the congested area to increase up to the level that compensates the marginal unit. However, it has encountered problems of its own. As fuel prices increased after 2002, expensive peaking units have tended to run less and less. As they did not achieve their 2002 capacity factor levels they did not fully recover their costs. FERC had assumed that since under LMP all generating units in a region would receive a price equal to the highest accepted PUSH bid, then it would be possible for a unit to receive a price greater than its own PUSH bid – and that the risk of under-recovery was therefore balanced by the risk of over-recovery. But this does not appear to have happened. Since these units often run at minimum levels when they do run, under ISO-New England pricing rules they are not deemed to be price-setting units, even if their PUSH bids are higher than the unit actually dispatched on the margin.

C. Local ancillary services markets and ''scarcity'' pricing

An important aspect of reliability is ancillary services and particularly operating reserves (i.e., generators that are either spinning or able to come on line within a specified time to prevent real-time imbalances between supply and demand). In markets with LMP, operating reserves should be jointly optimized and priced with the energy markets, and ancillary service markets should recognize local operating reserve requirements.[8] A co-optimized system of energy and operating reserves that

 1040-6190/$–see front matter , doi:/10.1016/j.tej.2006.12.006

recognizes local constraints can potentially provide more efficient market-clearing prices during reserve shortage hours and thus encourage the investments and operational decisions that make it more likely that a unit will be ready when it is needed. Such prices might, for example, encourage more investments in quick-start units or units with faster ramping rates in import-constrained areas. Unfortunately, since price caps are likely to limit energy prices in these same periods, they also blunt the incentives and the effectiveness of locational signals.

Local operating reserve markets affect another mechanism used in some electricity markets, the ''scarcity pricing mechanism.'' The NY-ISO uses an administrative scarcity price adder when the system as a whole is experiencing shortages of operating reserves (typically $1,000/MWh). When scarcity pricing is triggered the resulting high prices provide market revenue to owners of generation capacity that is operated during a limited number of hours of very high demand, reducing the need for RMR contracts. However, in order to be effective, it is essential that this mechanism is triggered by the physical realities modeled by the ISO both in the day-ahead and in real time. In other words, the dispatch algorithms used in the day-ahead and real-time markets need to recognize the locational reserve requirements that the ISO must adhere to when operating the system. The scarcity pricing mechanism would then raise prices in areas that do not fully meet local reliability requirements even when system capacity is sufficient to meet load and operating reserves in the region as a whole. However, as long as the scarcity price cap is still significantly lower than the value of lost load (VOLL), this mechanism can only complement, and not replace, a locational capacity payment mechanism.[9]

In practice, there has been a great deal of skepticism about the ability of FTRs to fulfill this role, or even to contribute to the process of capacity expansion.

D. Merchant transmission and FTRs

Transmission is an entirely different approach to ensuring capacity adequacy in constrained areas but can perform a reliability function equivalent to that performed by generation. When the underlying economic value of capacity differs between two regions, additional generation capacity may be warranted in the region with the higher capacity value. It is possible, however, that the most efficient solution would be new *transmission* between the regions so that a capacity-deficient region can benefit from the capacity surplus in an adjoining region. A better system of financial transmission rights (FTRs) could help bring about transmission solutions more regularly.

Most LMP energy markets issue FTRs which entitle their holders to a refund of congestion costs (the difference in LMPs at the source and sink of the transmission path). In PJM and New York, merchant transmission investors expanding the capacity of the transmission system can also be granted FTRs to provide incentives for decentralized transmission expansion. In practice, however, there has been a great deal of skepticism about the ability of FTRs to fulfill this role, or even to contribute to the process of capacity expansion. The problems include:

- Congestion rents drop dramatically when transmission is added;
- It is difficult to give property rights to merchant investors in a network when the construction of one piece of equipment in one part of the network alters transfer capabilities in many other parts;
- Central planners err on the side of caution so that centrally planned transmission investments end up being made before the FTRs get a chance to do their job; and most importantly,
- Price caps in the energy market may limit FTR values to below their economic value.

The price cap in the energy market has the effect of failing to signal the true value of additional transmission investment. For

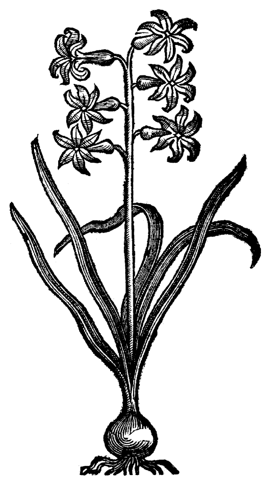
 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

example, in a region where all import capability is used to full capacity, and the region runs out of generating capacity, prices will be capped at the regulatory-imposed level (e.g., $1,000/MWh in the Northeast markets). Assuming the price of power in the exporting region is $100/MWh, because that region still has available low-cost generating capacity, then transmission is valued at $900/MWh in that hour. This may be well below the true value of transmission at the time, given the true value of lost load. While outage events are rare, it is precisely this value of mitigating high-value and rare events with transmission that should be signaled to potential transmission investors.

Electricity FTRs in the U.S. have traditionally been short-term (one-year duration), although FERC recently approved long-term FTRs.[10] However, the key challenge remains: how to better signal the true economic value of transmission to meet locational generating capacity requirements.

E. Locational access charges

Another alternative to help ensure adequate capacity in constrained areas is to send capacity price signals via the regulated transmission charges. Starting with the premise that both generators and customers pay for access to the grid (recovery of embedded transmission costs), specific locational access charges can be devised to encourage generators to locate in areas that will help to solve congestion problems. Annual access charges for generators in Ireland, for example, are currently based on the so-called MW-mile method. Based on a load-flow analysis, the model rewards generators producing reverse load flows and imposes charges on those that increase dominant flows. The



level of cost allocated to a generator depends on the distance and direction of resultant changes in power flows due to the existence of that generator.

A major disadvantage of this approach is that locational access charges are reassessed annually and the potential for significant change from year to year creates additional risk for generators. The approach also has short-term efficiency drawbacks in that the locational access charges reflect embedded rather than marginal transmission costs and market energy prices are not locational. The price in any given hour can be inconsistent with the marginal costs of efficient generator dispatch.[11] Essentially, locational access charges are an attempt to correct for the lack of locational energy prices and locational capacity price signals. Throwing such an attempted correction for the lack of locational capacity prices on top of an attempted correction for the lack of locational energy prices is at most a third-best solution.

A more appropriate time to give generators a locational capacity charge via transmission pricing signals is before the investment costs are sunk. ''Deep'' connection charges can effectively be set locationally to recover the cost of the transmission upgrade necessary to ensure that new generation capability – where ever it is built – is deliverable to the whole system. In contrast, the generally lower ''shallow'' connection charges do not send a price signal related to the cost of ensuring deliverability.

PJM (and other U.S.) deliverability requirements represent a form of deep connection charges for new generators. PJM requires new generating units to meet a ''deliverability standard,'' which requires adequate transmission service. Capacity resources must be deliverable to the total system load, including portions of the system that may have a capacity deficiency. If new generating capacity is not deliverable from its chosen location, it must pay for the necessary transmission enhancements to make it deliverable.

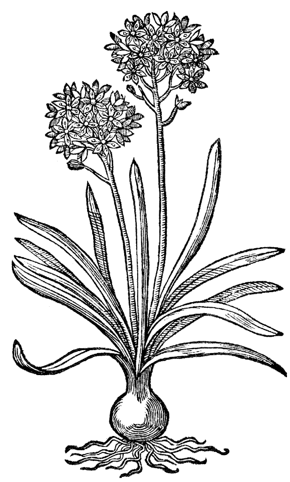
Deep connection charges presume that all generating capacity should be deliverable everywhere in the system on an equal basis. However, if under the most efficient allocation of resources the value of capacity is different in different locations, imposing deep connection charges would be inefficient. Total system costs could be lowered and the same reliability standards met without imposing this deliverability constraint.

F. The ''non-wires solutions'' procurement

Other non-CPM methodologies for helping to ensure capacity adequacy in constrained areas fall into the category of ''non-wires solutions'' (NWS), or ''transmission alternatives.'' NWS may include payments to new generation located within a constrained area, distributed generation, payments for committed demand reductions, or any network arrangement that allows grid investment to be deferred or avoided. In the U.S., NWS are often mentioned by state regulators as potentially useful mechanisms that should be part of regional transmission planning process. NWS initiatives have been particularly popular in BPA and the Northeast, which have a history of reliance on energy efficiency, demand reduction, and distributed resources.

The NWS approach serves as a centrally operated mechanism in which generation and transmission solutions compete against each other, with ideally the lowest-cost combination projects emerging from the process. The ISO or regulator determines the net market benefits of investments in transmission assets as compared to NWS, and the most appropriate choice in each location, taking into consideration grid reliability standards.

NWS has generally proved difficult to implement. New



administrative arrangements should be clear and economically sensible, and should be implemented in a way that does not distort markets that primarily rely on decentralized decision-making. However, like RMRs, the process by which non-wires generation solutions get ISO-sponsored payments may not be compatible with a competitive-marketplace.[12] In addition, the selection process is often arbitrary, mostly due to practical difficulties comparing transmission with generation and load alternatives as well as the choice of the time-frame assumed for the deferral of the transmission project. Transmission projects tend to provide more capacity than generation investments, and short deferral periods work against non-wires alternatives, particularly generation-based options.

III. Locational CPMs

A locational dimension can be introduced into CPM schemes (capacity requirements and/or payments) by defining them on a zonal basis to recognize that the true economic value of capacity in an import-constrained area might be greater than the regional average. The marginal capacity provided by plants or demand response resources in load pockets is more valuable than equivalent capacity outside the constrained area that cannot help prevent loss of load within the load pocket. The choice of a particular CPM design requires a case-by-case examination of the particular characteristics of the system and the stage of restructuring. In any case, the overarching objectives of any CPM are:

- Reliance on market-based principles, to the extent possible, as opposed to regulator-imposed solutions;
- Consistency with least-cost and reliable dispatch, providing efficient availability incentives for generators; and
- Provision of price signals that approximate the *true* economic value of capacity resources by time period and by location to guide required investments for reliability.

 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

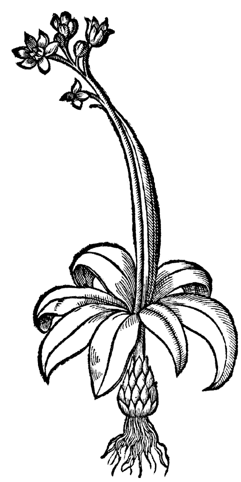
The relative merits of a locational CPM must be determining upon a careful analysis of the likely impact on a number of key areas:

- What efficiency distortions are introduced in the market (energy prices and quantities, level of reliability and cost) by the specific elements of the capacity payment construct?
- How effectively can the locational capacity payment work to solve ''load-pocket'' constraints?
- Can transmission and generation solutions compete under the approach?
- Do locational compensatory schemes alleviate or increase market power concerns?

The best CPM solution will be the one that is more likely to achieve capacity adequacy without sacrificing other goals of the overall capacity solution, including any energy policy goals such as limited volatility in capacity prices. In principle, locational CPMs in electricity markets may take three possible forms: (a) a regulated (fixed) payment made to all available capacity (price-based CPMs); (b) a regulatory-mandated capacity quantity imposed on all load-serving entities of the system (quantity-based CPMs), or (c) a hybrid scheme where the capacity price is set by a formula that links price to actual capacity levels (hybrid CPMs).

One major drawback of price-based CPMs is that when the price is fixed the quantity of capacity delivered by the market can be highly uncertain, and such uncertainty undermines the reliability objective, one of the main policy reasons for creating a CPM.

U.S. capacity markets have traditionally adopted the quantity-based approach. The ISO creates demand for capacity by imposing a minimum capacity requirement for the region or pool, and each load-serving entity (LSE) is



required to provide a share of the regional reserve requirements, generally based on its previous year's contribution to system peak. The main disadavantage of quantity-based approaches is the extreme volatility of their price outcomes, with capacity prices fluctuating from $0 in surplus periods (because there is no demand for incremental capacity) to the deficiency charge or administrative penalty in periods of shortfall.[13] This high volatility creates risk for both consumers and investors, which ultimately increases the cost of capacity or the cost of meeting a particular capacity target.

A ''hybrid'' CPM approach was used in the previous England and Wales Pool,[14] and was pioneered in the U.S. by the New York ISO. The advantage of a hybrid approach is that it can be structured to limit the price volatility of quantity-based CPMs and reduce the quantity uncertainty of the fixed-price CPM. The parameters are set to achieve a target level of installed capacity and LSEs purchase more or less capacity relative to the target depending on the capacity price. Hybrid CPMs may be used to implicitly set a capacity price cap to limit market power. A disadvantage of hybrid designs is that they are likely to be particularly sensitive to regulatory intervention, and subsequently they tend to increase the perceived regulatory risk.

A. Locational capacity demand curves

Some locational hybrid CPM schemes in the U.S. rely on ISO-administered capacity demand curves. All generators that located within a capacity zone that participate in the ISO auction are entitled to the resulting zonal capacity price. The auction uses a downward-sloping demand curve, whose price parameters are set by the ISO, not the actual buyers' bids. The ''markets'' are cleared for all zones at the same time and all LSEs are charged the local clearing price for the capacity purchased on their behalf. The auction results in higher prices in the transmission-constrained zones, which provide

 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

more revenues to generation owners in those areas. Consumers in the constrained areas bear the extra cost. New York currently operates a capacity demand curve, with monthly auctions that clear the market in three distinct zones: two ''load pockets'' (New York City and Long Island) and the rest of New York State.[15] PJM's proposal (Reliability Pricing Model – RPM), also includes a regulated demand capacity curve and establishes different prices by capacity zone (Locational Delivery Areas).[16]

In 2004, the New England ISO proposed a Locational Installed Capacity Payment (LICAP), aimed at reducing the ISO's increasing reliance on out-of-market reliability contracts. LICAP would involve separate capacity auctions for three import-constrained zones (Northeast Massachusetts, Southwest Connecticut, and the remainder of Connecticut), one export-constrained zone (Maine) and the remainder of the New England pool (''Rest of Pool''). FERC initially approved the LICAP proposal, but it was later abandoned due to strenuous opposition by state regulators and customers who argued that it was a costly administrative approach that would not provide assurance of new capacity in load pockets.

The goal of any administrative capacity demand curve is to minimize the probability that installed capacity falls below a minimum level, its ''objective capability.'' The New England ISO demand curve was designed to provide generators with higher price signals when actual capacity was near that minimum (112 percent of expected peak load). When actual capacity levels fell below the Objective Capability, the curve assumed that consumers would be willing to pay up to two times the cost of new entry (''Efficient Benchmark Capacity Cost,'' or EBCC).[17] The clearing price would be determined by the point of the demand curve corresponding with the total installed generation capacity in the zone (regardless of whether some of those units were temporarily de-listed), and adjusting for exports and imports to and from other zones (**Figure 1**).

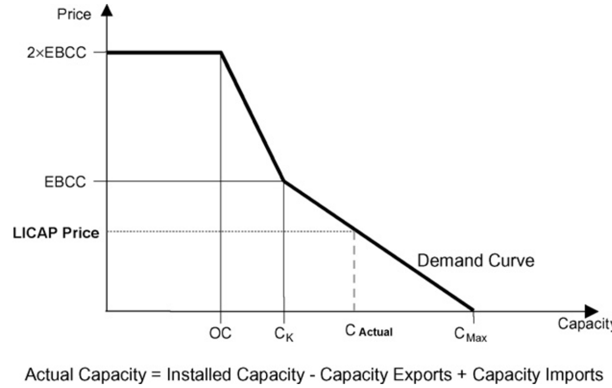


Figure 1: ISO New England's LICAP

There are a number of problems with a capacity demand curve of the type proposed in New England.

The first problem has to do with the short-term nature of the capacity obligation. Since the capacity offered and purchased is based on monthly markets, there are no price signals for long-term investment, and investors may perceive revenue streams as too uncertain.

The second problem has to do with the reliability levels that would obtain remuneration under such a scheme. Economically speaking, the efficient level of capacity should reflect consumers' VOLL. Under the LICAP demand curve, there would be a positive capacity price above the target capacity level, up to about 128 percent of the peak load. Payments could be made for capacity that are greater than their value of the incremental reserve to the consumer. This potential overpayment in the NE-ISO proposal was, in part, a consequence of the ISO's uncertainty about the level of capacity that the demand curve will encourage.

Third, an administrative demand curve is a heavily regulated approach. Excessive regulatory uncertainty can increase the cost of capital of generators and reduce investment below the level desired by the regulator. The main argument in favor of a demand curve approach is that it reduces price volatility, which allows developers to better predict their ICAP revenues and finance new

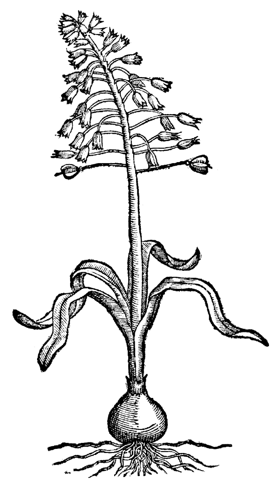
 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

capacity. However, the regulatory nature of the curve might work in the opposite direction and increase risk associated with plant investment (e.g., the regulator might just redefine the slope of the curve to reduce the payments to generators and the clamoring of consumers).

The New York ISO ICAP demand curve, in operation since 2002, appears to have met its immediate objectives of reducing price volatility and improving the transparency of the ICAP markets. However, it remains to be seen whether the ICAP demand curve with monthly capacity obligations will meet its longer-term objective of encouraging sufficient capacity region-wide, at a reasonable cost.

B. Locational forward capacity auctions

Typically, quantity-based CPMs have been designed with one-year time horizons. A novel approach within the group of quantity-based CPM schemes is ISO-operated capacity auctions that use multi-year commitment periods. The problem of price volatility fades away with a multi-year capacity auction. Such auction schemes can be designed to accommodate locational constraints. Under these schemes, the ISO is still responsible for performing forecasts of the amount of capacity (ICAP) required in the region (or locations within the region) but it does so for three or four years in the future. If the auction is designed so that it occurs with sufficient time in advance of the actual performance date or commitment period (say, three to four years) both existing and potential new entrants will be able to compete for the same product and the price will reflect long-term (improved) surplus conditions.[18] This has three important positive effects:



- Expanding the pool of competitors to include entrants reduces market power concerns. The increased competitiveness can lead to lower capacity costs and generators' expectations of revenue in energy and ancillary service markets being reflected in their offers.
- The three-year lead promotes stability of capacity prices, which will tend to oscillate between the fixed O&M costs of existing plants (in periods of surplus capacity) to the long-run marginal cost of capacity. If new capacity is required, the auction capacity price will rise to the level required to attract entry. The capacity clearing price should therefore vary around the competitively derived (incremental) cost of new entry (CONE).
- It encourages innovation – existing generators compete with current and future resources, which may encourage them to upgrade or invest in new technologies by the time the performance is required.

A multi-year forward capacity auction may be combined with administrative demand capacity curves, as illustrated by PJM's Reliability Pricing Model (RPM) proposal, where PJM would run auctions for capacity products to be delivered four years later, using a capacity demand curve. Forward Capacity Auctions that rely on demand capacity curves still retain the regulatory risk associated with a heavily regulated approach. A particular form of capacity auction that employs a *market* rather than an administrative process for capacity procurement, while also accommodating congestion problems, is the ''Descending-Clock Auction (DCA)''.[19]

1. How forward ''descending-clock auction'' schemes work

Under a DCA scheme, each year the ISO runs a descending-clock auction to find enough resources to meet the region's ICAP in the commitment year (e.g., resources needed in the summer of 2012 will be auctioned in the spring of 2009). The auction begins with the ISO naming a price, and suppliers indicate the quantity of capacity they are

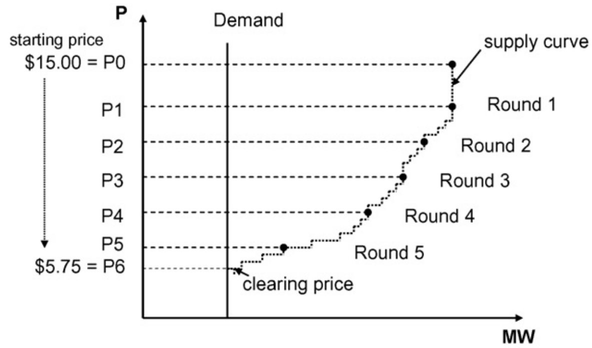


Figure 2: Illustration of Descending-Clock Auction

willing to offer at that price. If there is more capacity supply than needed, the auctioneer lowers the going price for the following round – the ''clock ticks down.'' When the total amount of MWs submitted is just as large as the capacity needed to meet the required ICAP, the auction closes. The winners are the bidders in the last round of the auction. They will receive the auction winning price for each kW of capacity committed.

Figure 2 below illustrates an example of a descending-clock auction structure.

The clock auction can accommodate multiple zones or multiple products by running several simultaneous clocks that tick down at different speeds, depending on the excess supply for each of them. For example, if there is excess supply in zone A but not in zone B, the going price for zone A will be lowered in the subsequent round, but the price for zone B will stay the same. Bidders are still not allowed to bid more in a round than they bid in a previous round but this restriction is only valid for the whole auction and not for each zone. Subject to not creating aggregate excess demand, bidders are allowed to switch their bids from zone A to zone B and then switch back in a later round. In this way, the market will determine the relative prices between A and B and the entrant can make an informed decision as to where to build the plant. The auction closes when there is no excess supply in either zone.

We see numerous advantages to this market design.

- A capacity DCA potentially requires less regulatory intervention than a LICAP-demand curve mechanism, therefore lowering regulatory risk. It can also accommodate retail load switching for LSEs.
- The dynamic character of zonal designations may enable state regulators and load to identify and take actions to eliminate transmission constraints creating capacity zones. Therefore, the auction may provide a useful locational economic signal to all market participants to solve constraints and locate where needed.
- A descending-clock auction format (in contrast to simultaneous sealed-bid uniform price auctions) can provide a more efficient outcome, as it can better address the decision-making problem that bidders face when they enter the auction. The auction process ensures that the ISO purchases capacity from the lowest-opportunity-cost resources (the bidders in the last round). The more efficient, least-cost resource providers can simply wait for the less efficient resource provider to withdraw from the auction.
- Finally, a DCA approach presents advantages over adequacy methods that rely on bilateral contracting because the bidding process reveals crucial information. In an open-auction format, bidders learn valuable information about other bidders' estimates as the auction proceeds and bidders drop out. Such information may be used by a bidder to modify its expectations of energy and ancillary service market revenues as well as of system excess supply at a given price level.

2. *Illustrating a locational DCA: the New England FCM*

The Forward Capacity Market (FCM) adopted in New England illustrates this type of capacity auction and can be used to discuss the impact of this promising capacity approach. It was officially approved by FERC on June 15, 2006. The first auction is expected to be held early in 2008 for a commitment period beginning June 1, 2010. The starting price of the auction will be two times the agreed starting value of cost of new entry of

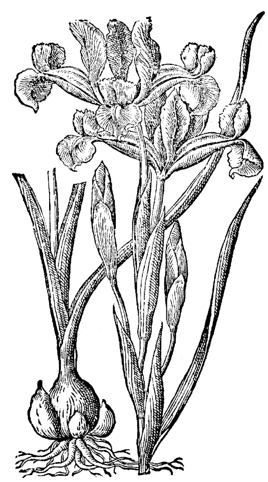
$7.50/kW-month or $15/kW-month.

The FCM contains a locational component that allows prices to differ between import- and export-constrained zones within New England. The need for a locational component will be evaluated by the ISO before each auction. If transmission limits (including transmission upgrades that are predicted to be on-line by the commitment period) are expected to bind, capacity zones are designated by the ISO and separate but simultaneous auctions are held for each zone to meet a Local Sourcing Requirement. Potentially the auction will stop the clock at a higher price in an import-constrained zone or at a lower price in an export-constrained zone. Each year the ISO would determine a zone's capacity requirements in proportion to its share of the NE Control Area's coincident peak in the previous year.

The FCM also includes a locational element in that capacity payments will recognize the potential capacity element in the hourly LMP prices. Specifically, payments for capacity will be reduced by an amount known as the peak energy rent (PER) when the relevant LMP exceeds a specific strike price equal to the deemed incremental cost of a marginal proxy unit. The PER will be computed as the revenues from a 12-month rolling average of LMP less the variable costs of a hypothetical, benchmark combustion turbine unit.[20] Since the PER will be linked to energy LMPs, given transmission constraints, PER will vary by location.

Other elements of the FCM include:

- *Commitment period*: Existing generators accepted in the auction will only commit capacity resources for one year at a time,



while new capacity will be able to lock in a compensation at the market-clearing capacity price for up to five years. The goal is to provide a predictable return on the investment for the initial years of operation.[21]

- *Availability conditions*: Generators would be paid for any capacity purchased from them, but not if the capacity is unavailable when needed.

On any critical day, a resource can have its compensation reduced up to 10 percent of annual FCM payment if it is not available during shortage events, and, in any month, a resource can lose up to two and one-half months of its annual FCM payment. The forfeited capacity payments will be distributed among the generators that actually performed during the shortage events.

- *Reconfiguration auctions*: The ISO will conduct annual reconfiguration auctions one and two years prior to the commitment period to allow suppliers who had previously been selected in an FCM to exchange that obligation with other suppliers. Seasonal and monthly reconfiguration auctions will also be held during the commitment period. The ISO can use these reconfiguration auctions to buy additional capacity resources or sell back resources it may no longer need. Reconfiguration auctions mitigate risk to resource providers as their supply commitments come due, especially the risk of large deficiency charges.

IV. Challenges of Locational CPMs: Looking Ahead

Notwithstanding the advantages of DCA-based auctions, there are challenges ahead as to how these schemes will effectively work. Four areas in particular will need further review to increase the probability that CPM schemes can succeed in restoring the missing locational signals.

A. Will locational incentives work?

Locational forward capacity schemes such as the one approved for New England take into

 1040-6190/$–see front matter , doi:/10.1016/j.tej.2006.12.006

account transmission upgrade proposals expected to be operative by the start of the commitment period. These transmission proposals will be an important factor in final auction capacity prices by zone but it is unclear how incentives for merchant transmission will work. Both the ISO-NE and PJM proposals use capacity transfer rights (CTRs) to hedge congestion costs (differences between zonal capacity prices) for each interface. The details of the CTR allocation process have not yet been established, but incentives to invest in transmission upgrades through CTRs may be limited – the same problems that limit the effectiveness of FTRs associated with capped energy LMPs may also apply to merchant transmission. Centralized transmission planning decisions compensated at regulated, postage stamp rates will still have a major impact on the effectiveness of locational capacity payments in solving locational problems.

It is essential that these transmission projects are not mere speculations. If approved transmission projects suffer delay, the locational price signals will not reflect the actual needs in an import-constrained zone in the year resources were committed. A system of penalties for failing to complete the promised project would need to be devised (perhaps, by allocating the extra capacity costs arising from reconfiguration auctions to those parties who failed to build).

At the same time, if new entrants accepted in the zonal capacity auction are aware that a transmission owner (TO) is planning to build transmission lines sometime after the three year lead-time period, they will also be aware that the completion of the transmission project will depress

the zone's capacity price. If generators are to internalize expectations of future zonal capacity prices, the system operator must make sure that these plans are transparent. It should also give the opportunity for other solutions to be proposed in the initial stages of ISO or TO's planning process.

Another anticipated source of competition in local markets would be the participation of demand-response in the multi-year capacity auctions. However, it is still to be seen how demand-response participation would work in the capacity auctions, especially as customers are generally unwilling to commit to demand reductions several years in advance.

B. Definition of capacity zones

Locational CPMs can help market participants form expectations of congestion costs by zone. This information may be useful to LSEs in signing long-term bilateral contracts to hedge capacity costs corresponding to the area where they are located. However, setting the criteria and defining the zones is not a straightforward matter and the choices affect both reliability and the transfer of risk among parties. A potential problem with forward auctions and locational schemes is that the actual value to the system of resources committed through long-term contracts would be affected by any changes in capacity zones from year to year, leaving parties exposed at the point of the commitment period to the risk that the locational capacity price will change. Unless a very transparent and objective process is devised the risk is even higher if the definition of capacity zones is subject to regulatory proceedings, as is the case in New England, where the stakeholder process provides for consultation with state utility regulatory agencies. Any uncertainty arising from lack of objective processes may deter local bilateral contracting by the LSEs.

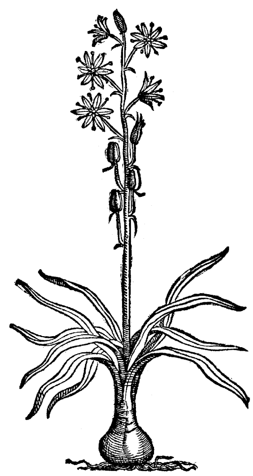
Finally, constraints could warrant different energy prices by

 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

node within a zone so that the broader zonal prices can create incentives for investment and operational decisions that are inconsistent with the reliability needs. Whether the zonal capacity auction eliminates the need for additional ''out-of-market'' compensation schemes for selected local generators important for reliability will depend on how successfully the ISO draws the capacity zonal boundaries and how accurately they reflect the actual constraints that limit the deliverability of power on the grid. Other locational considerations include the need to coordinate with any locational forward operating reserve (spinning reserve) markets and with interconnection policy – e.g., with a policy of ''deep'' interconnection charges, the generator internalizes the cost of reinforcement in its capacity bid.

C. Local market power

Distinguishing between prices that reflect true scarcity conditions and prices that result from the exercise of market power (capacity withholding) is difficult and contentious. In the FCM design, incumbent generators will have little opportunity to affect price. An FCM auction will control market power by subjecting existing generators to competition from the potential new power plants and merchant transmission, whose bids will set the auction clearing capacity price. In addition, the market monitor will review offers by existing capacity suppliers for signs of the exercise of market power and makes the prices public so that developers can estimate how much new capacity will be needed and propose new projects to compete with existing capacity resources.



While in principle local market power mitigation should follow the same methodology employed for system-wide market power, the situation is particularly tricky in small markets. The key to a locational capacity scheme is to define zones broadly enough so that competition can work, and generating units can be scaled efficiently. In practice, if a potential new efficient unit is oversized for the load pocket, there is a tradeoff between having a more costly, smaller-scale new unit to fit the peak load growth (leading to higher local capacity prices in the local area) and carrying out transmission upgrades so that larger units can export to outside areas.

Further, controlling market power by putting incumbent generators at the mercy of the price-setting actions of new generators only works if the new entrants are willing to act as price makers. However, a new generator with a guaranteed contract from some other entity might have very little incentive to act as a price maker in the FCM, and be content rather to be a price taker. If too many new units are price takers, there may be few bidders in the FCM and the market could collapse. The potential for this problem is exacerbated if not very many new units are needed in the first place. In small geographic markets like New England, which only grows about 600 MW a year, it is likely that new units will be added infrequently, so that the ability of the system to contain market power through the bids of new entrants is limited.

D. Link to energy markets

In equilibrium, the clearing price in the capacity auction should be expected to equal the capital costs of an efficient new entrant peaker less the rents that such peaker would expect to earn in the energy market (the PER). In the presence of an explicit PER adjustment mechanism, generators will internalize the expected PER reduction when submitting bids in the

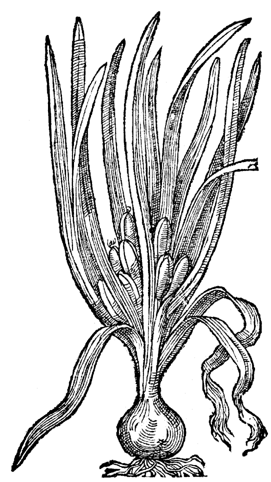
 1040-6190/$–see front matter © 2007 Published by Elsevier Inc., doi:/10.1016/j.tej.2006.12.006

capacity auction. In order to avoid jeopardizing the intended goals of the locational CPM, the PER, as a minimum, should meet the condition that real-world units are capable of remaining as efficient as the benchmark generator over time. If there is a new, more-efficient technology built, revenues to existing less-efficient units will be lower, reflecting the fact that they are now less valuable.

Depending on the specific design, PER approaches may have potential negative implications in the efficiency of the energy markets. When a PER adjustment is established so that it effectively acts as a contract for difference (CfD) – i.e. so that any difference between the spot price and the strike price is effectively deducted from energy market revenues to avoid double payment – there is a risk that generators will not have incentives to offer energy prices above the strike price. In the extreme, one could expect the spot energy price at or below the strike price in practically all hours. This outcome would create poor short-term incentives to be available at times of scarcity in the energy market. Availability incentives would largely rely on administrative decisions (penalty-avoidance incentives) at the expense of market-based signals of scarcity. The FCM's 12-month rolling averaging approach adopted in New England may mitigate this effect somewhat but it nevertheless deserves some careful analysis.

V. Conclusions

The ''first best'' solution to the existing electricity market failures requires enabling the demand side to work along with regulatory willingness to let



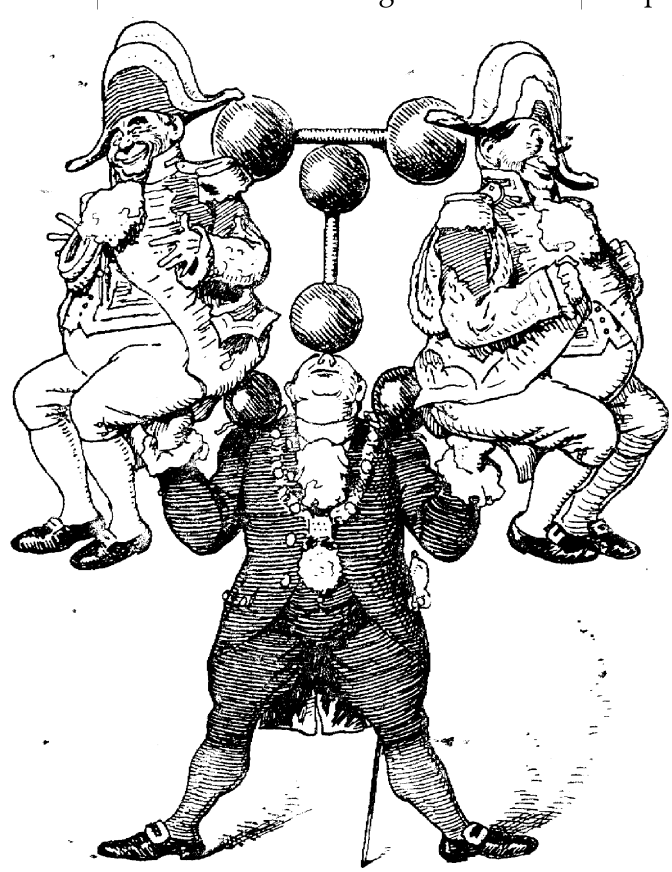
market prices rise when the marginal cost of supply increases. Unfortunately, this is not the case in many markets. The current price caps in the energy markets send distorted price signals and limit incentives for long-term bilateral contracting. Until the conditions for a first best solution are in place, public policy and market designers must continue looking for mechanisms to approximate the true economic value (time- and location-differentiated) of capacity, while at the same time reducing concerns over price volatility and market power. In this article we have focused on the locational aspect of the value of capacity. First, we have reviewed a range of methods to solve locational problems outside of a centralized CPM construct, and showed how these methods either present problems in terms of efficiency goals, or are otherwise limited in their success to provide the required incentives. We then reviewed possible locational ISO-operated CPM designs.

Locational CPMs are not a perfect substitute for LMP energy-only markets. First, they do not provide real-time hourly marginal capacity cost signals as would an energy-only market. Second, the CPM reliability target is still generally chosen by the ISO or a central authority, instead of being driven by the level of reliability that consumers are willing to pay for. Nevertheless, well-designed CPMs hold more promise for signaling the value of capacity than the non-CPM alternatives. Locational CPMs that rely on market-based principles, such as forward capacity auctions, are superior to cost-based payments directed to specific must-run generators, as CPMs at least provide a meaningful price signal about the opportunity cost (or economic value) of resources to potential investors. Similarly, while capacity demand curves often translate into purely administrative capacity prices, forward capacity auctions with a three- or four-year lead time period effectively allow the market (the new entrant) to set the capacity price.

 1040-6190/$–see front matter doi:/10.1016/j.tej.2006.12.006

With the proper design, capacity price differentials between zones will be expected to improve incentives to build required generation and/or transmission infrastructure where required. The rules that determine the generators' entitlement to capacity payments should ensure that the peaking capacity is there when and where it is needed. If the rules are effective, the underlying expected loss of load probability during high peak hours should not be systematically higher than in a first-best solution scenario. Finally, locational CPMs may redistribute costs to end users in congested areas, and eventually, they may lead to regulatory reserve margin requirements being replaced by price-responsive, voluntary curtailable load, especially in congested areas.

Locational forward capacity markets are very new and therefore their impact on providing locational solutions and overall resource adequacy is yet to be seen. Market participants will need to gain confidence in the new design and the stability of these programs will be an important factor in ensuring that generators can rely on the CPM as a source of adequate returns on new investment. To the extent that locational CPMs are able to provide stable capacity prices while preserving long-term system reliability, they are a step in the right direction. If the New England FCM proves successful, other jurisdictions should consider introducing a locational dimension to CPMs, with the caveat that no locational CPM approach can be regarded as the ''one-size-fits-all'' solution. The choice of the right locational capacity solution involves an analysis of the specific characteristics of the sector as well as important decisions internal to the scheme. Finally, locational CPMs combined with innovative retail pricing options that reflect the expected higher costs at times of peak demand in congested areas will further encourage demand-side participation when and where it is needed. The goal of all this should be that regulatory reserve margin requirements will be eventually replaced by economic reliability levels, as signaled by price-responsive loads.■



Locational forward capacity markets are very new and therefore their impact is yet to be seen.

 1040-6190/$–see front matter © 2007 Published by Elsevier Inc., doi:/10.1016/j.tej.2006.12.006

Endnotes:

1. Though recently, there has been increasing focus worldwide on deploying interval metering and implementing dynamic rate programs on a large scale. For further discussion on this subject see *Responding to EPAct 2005: Looking at Smart Meters for Electricity, Time-Based Rate Structures, and Net Metering*. EEI, by Ken Gordon, Wayne Olson and Amparo D. Nieto, May 2006.

2. In the US, for example, generation bids are capped at $1,000 per MWh in New York, New England, and PJM, or $400/MWh in California.

3. Even in the context of uncapped markets, the uncertainty in the frequency with which prices spike means that generators may decide against building new capacity unless they can rely on long-term contracts to cover their fixed costs. (Spot market revenues alone may not form a sound basis from which to plan for generation investment, unless the expected economic profit is substantial.)

4. Performance-Based-Regulation (or PBR) approaches are not part of the scope of this paper. In cases where non-profit ISOs and RTOs plan and operate the transmission system, the focus is necessarily shifted toward providing incentives to market participants.

5. They can also be applied in load pockets that are not workably competitive.

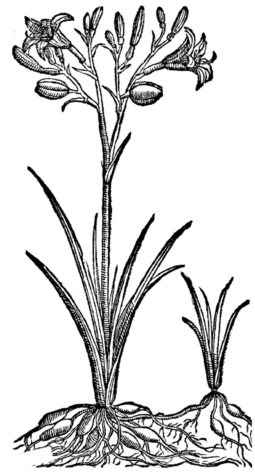
6. This is the case, for example, in New England. In PJM, the extra cost associated with RMR payments is charged to the specific local load zones.

7. PJM ISO recognized this need and developed rules that would allow a unit to defer a retirement *for a limited period*. The ISO builds the transmission to operate with the retirement and RMR payments only apply until the transmission is complete.

8. ISO-NE, for example, is planning to implement co-optimized locational energy and reserves real-time markets.

9. A VOLL as high as $10,000 per MWh is often used in planning studies to set installed reserve requirements. Studies to survey the VOLL show that the value varies by type of customer. However, at the margin, the VOLL that would set the market price in an hour of insufficient capacity would be the highest accepted demand bid in the market.

10. Order 681 requires RTOs to make available long term TFRs that assure rights for a minimum of 10 years. FERC left open the question of whether the 10-year minimum could be satisfied with shorter duration



terms that include assured renewals for at least 10 consecutive years.

11. The new pool-based market proposed for the island of Ireland will therefore require constraint payments, payments made to those generating units that were scheduled in an Ex Post Unconstrained Schedule (EPUS) but did not actually run because of transmission constraints.

12. For example, when the CAISO attempted a NWS project in 2001, a number of stakeholders saw this process as a return to a central planning regime, and claimed that such decisions should be left to the market players.

13. Under most RAR rules, if a LSE does not satisfy its resource requirement, it is assessed a deficiency penalty, which sets a 'de facto' capacity price cap.

14. In the price formulae used in the E&W power pool, an hourly LOLP element produced additional revenue in an inverse relationship to the system reserve margin by hour.

15. An LSE in New York City must procure 85 percent of its ICAP from in-city generators. In Long Island, 93 percent of the Island's peak load must be supplied by local capacity. A generator located in another jurisdiction but committed to the load pocket via an AC or a DC line also qualifies towards the locational requirement.

16. In addition, PJM-ISO would use deliverability requirements to ensure that a generating unit is actually providing a value to the system when it is needed. The PJM proposal would have two zones initially and more zones (up to 23) later.

17. The NYISO's demand capacity curve is also capped at 2 × EBCC, but it reaches that maximum price at a lower level of reserve margin.

18. As is the case in most of the CPMs in the US, LSEs can still self-supply through their own resources or bilateral contracts, which will be taken into account by the ISO to offset the LSE's projected capacity-obligation. Reductions in demand through energy efficiency and dispatchable demand response programs would also qualify as capacity.

19. A CPM that makes use of the DCA form was submitted to FERC on March 6, 2006 by New England generators, ISO-New England (ISO-NE), and four out of six New England states. NERA Economic Consulting reviewed DCA proposals and provided recommendations to ISO-NE, ISO-NY and PJM in 2003 and 2004.

20. The current NY-ISO ICAP market also includes a PER adjustment, but it is based on a *forecast* of PER energy and ancillary services rents over a year, assuming long-run equilibrium (i.e. assuming that installed capacity reserves equal the reserve requirement).

21. A similar CPM feature was adopted in Russia in late 2006. The CPM will include locational auctions with a three-year lead time and a 10-year contract term. The first contract is expected to apply from 2010.