Electricity Journal, 18, 43-54, August/September 2005.

A Capacity Market that Makes Sense

Peter Cramton and Steven Stoft[1]

15 June 2005

Abstract

We argue that a capacity market is needed in most restructured electricity markets, and present a design that avoids problems found in the early capacity markets. The proposed market only rewards capacity that contributes to reliability as demonstrated by its performance during hours in which there is a shortage of operating reserves. The capacity price responds to market conditions, increasing when and where capacity is scarce and decreasing to zero when and where it is plentiful. Market power in the capacity market is addressed by basing the capacity price on actual capacity, rather than bid capacity, so generators cannot increase the capacity price by withholding supply. Actual peak energy rents (the short-run energy and reserve profits of a benchmark peaking unit) are subtracted from the capacity price. This allows the capacity market to more accurately control short-run profits and suppresses market power in the energy market. This design both avoids and hedges energy market risk, and by suppressing market power avoids regulatory risk. Risk reduction saves consumers money as do the performance and investment incentives inherent in the pay-for-performance mechanism.

Capacity markets have proven to be one of the most contentious elements of electricity restructuring. Many argue there is no need for a capacity market. Others argue that, while they may be needed, the current designs are woefully inadequate, and are dominated by having no market at all (Cramton 1999) or alternative instruments (Chao and Wilson 2004; Oren 2004). Still others argue capacity markets are essential for encouraging sufficient investment in new capacity (Besser, Farr and Tierney 2002; Stoft 2002). We argue that a capacity market is needed in most restructured electricity markets, and present a design that avoids the many problems found in the early capacity markets. ISO New England (ISO-NE) has proposed a design for the New England market that relies on the principles we present here.

Fundamentally, the capacity market we propose is an incentive mechanism to induce supply to invest in sufficient generation in the right locations and of the right type to satisfy a reliability standard at least cost. The capacity market provides strong incentives for suppliers to perform when most needed, reduces risk for both generators and load, and addresses market power both in the capacity market and in the spot energy market.

[1] Cramton@umd.edu; Steven@Stoft.com. We are grateful to ISO-NE for asking us to think on this topic and especially to its staff and board for many helpful discussions. The views expressed are our own.

Why have a capacity market?

Most markets do not need to price capacity to promote efficient long-run investment. Pricing the primary good is sufficient. The balance of supply and demand in the spot market determines a clearing price, which in turn determines short-run profits for capacity. In the long-run equilibrium, capacity enters until the point where the expected short-run profit is equal to the carrying costs of marginal capacity. As supply tightens or demand expands, short-run profits increase, and this sends a "build" signal to suppliers.

Current electricity markets have an important market failure—the absence of a robust demand side—which motivates the need for a capacity market. Today, there is little demand response to the energy price, primarily because most load neither sees nor pays the real-time price. Real-time meters and demand management control systems are not yet in place for most electricity consumers. This absence prevents load's willingness to curtail demand to set the price during times of supply scarcity. Furthermore, the market structure is imperfectly competitive, especially in load pockets. As a result, there are instances when one or more suppliers has substantial market power, especially at peak times or during an outage of a large generator or transmission line.

These market failures require that (1) prices during shortage hours—those hours in which energy and operating reserve requirements are not fully met—be set administratively, and (2) rules be in place to monitor and mitigate bids in situations where market power is likely. Unfortunately, addressing these market failures typically results in price peaks that are too infrequent and short to motivate efficient investment in new capacity. Theoretically, it would be possible to set shortage prices sufficiently high to provide sufficient investment incentives, as is the case with value of lost load (VOLL) pricing (see Stoft 2002), but this approach entails estimating VOLL—a nearly impossible task. Moreover, it exposes load to greater price risk in real time, and the high shortage prices encourage generators to withhold supply to create shortages, which undermines reliability. Thus, in practice, shortage prices have generally been set at levels on the order of $1000, and even lower, such as the $250 cap in the California market during a time of great scarcity (2000-2001). In contrast, the value of lost load estimates are often in the $10,000 to $20,000 range.

A capacity market can supplement the revenues a generator can get from the energy and reserves markets. This capacity revenue allows regulators to set shortage prices at politically acceptable low levels, and yet generators still can be motivated to make efficient investment and operating decisions. However, for this to be the case, the capacity market must be carefully designed.

A complementary alternative view of the need for capacity markets is gained from the reliability perspective. Besides high shortage prices and capacity markets, other approaches to inducing sufficient investment for reliability have been proposed, among them enforcement of a high level of long-term contracting and increased demand response. When evaluating all of these proposals it should be kept in mind that no purely market-based solution to the reliability-investment problem is possible until consumers

can purchase reliability and this is impossible until a sufficiently large set of consumers can be individually and controllably disconnected during a supply shortage.[2]

Until individualized customer reliability is possible, all approaches except a very responsive demand side require at least one crucial administrative input. Shortage pricing requires an estimate of the VOLL. A capacity market requires an estimate of the reliable capacity level. And long-term contracting requires specification of the mandatory energy option contracts, the penalty for a failure to hold sufficient options, and the penalty for failure to perform when called. Notice that at least in shortage situations, the penalty for nonperformance needs to be set administratively at a level above the spot energy price to provide sufficient incentive for investment. Otherwise, the option price would be bounded above by the cost of nonperforming assets, which face a penalty equal to the capped and mitigated spot energy price minus the strike price whenever called to supply energy.

The demand response approach is in a different category because it does not induce a reliable level of investment. Instead it makes a range of investment levels reliable and induces the least-production-cost level of investment. Reliability is achieved by demand responding almost instantly to price rather than through operating reserves—otherwise it would still be the administrative control of operating reserves that determined reliability and not the market.

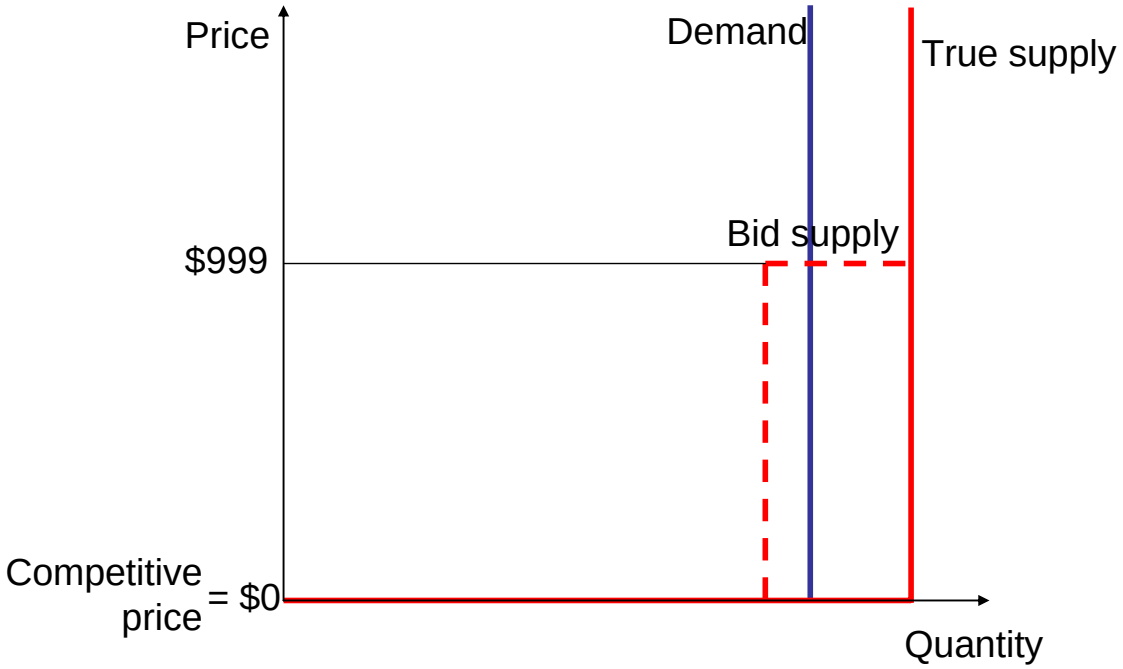
Early markets were fatally flawed

Early capacity markets did not come close to an efficient design. In their worst form, the markets simply were a means of transferring an arbitrary amount of money from load to generation, without load getting anything of value for the purchase. The markets were simple and seemingly sensible. Every month, generators would offer their capacity in a uniform price auction. The system operator (ISO) would accept the offers—lowest price first—until the required quantity of capacity was procured. All generators accepted were paid the market clearing price for each kW-month of capacity they offered.

Market power. A major problem with this design stems from the vertical demand curve and the short-term offer. For capacity offered on a monthly basis, the opportunities are limited, and hence the true marginal cost of this capacity is near zero. Thus, the competitive outcome is determined from the intersection of the true supply curve (near zero up to the system capacity and then infinite) and the demand curve (vertical at the capacity requirement), as shown in Figure 1. This yields a competitive price near zero, whenever the system capacity is sufficient to satisfy the requirement, which through planning should be all the time. This near-zero price gives large suppliers a strong incentive to exercise market power by bidding their supply at high prices, as shown below. A supplier is much better off supplying a reduced quantity at a high price than its full quantity at a price near zero.

[2] It is not sufficient that distribution companies be individually interruptible because their distaste for blackouts is not intrinsically determined but is instead determined by an administered penalty.

Figure 1. Competitive and uncompetitive outcomes in a traditional capacity market



The experience with this market suggests that, although it may be possible to sustain something close to the competitive price much of the time, suppliers do eventually figure out that higher bids are more profitable. The result is that the price is determined by the willingness of suppliers to exercise market power, rather than the efficient interaction between true demand and supply.

Product measurement. A second basic problem is the measurement of supply. All of the ICAP markets in the East Coast rewarded capacity based on availability—a unit that is 90% available gets 90% of the ICAP price. This is sensible, but the measurement of availability in all these markets was and is poor. Namely, the resource is paid for ICAP based on average availability over non-maintenance hours. This is done with the use of an engineering formula developed over the past thirty years for an entirely different purpose. It measures not how well a unit performs, but how well it performs relative to expectations for a unit of this type—and even then loopholes are provided. This measure is called EFOR'd and it is convenient because the data is collected anyway, but it is misguided and gameable.

To see the folly of this approach, imagine a "dog" plant with high marginal cost that is extremely difficult to get running. It would never want to be called for energy, because if it were called it would not be able to deliver; hence, it would bid into the energy market at an extremely high offer price ($1000) with a high start-up cost and a long (12 hour) start time, and a long minimum run time. Such a unit would never be called for energy. Even in crisis, when the ISO would gladly pay $1000 per MWh, the unit is not called, because the ISO believes that the crisis would have passed before this unit can get online. In spite of this, the unit will have a high EFOR'd that depends on its forced outage rate and it rarely breaks down when running. The unit receives the same ICAP payment as well-maintained baseload or peaking units that are often providing energy or reserves,.

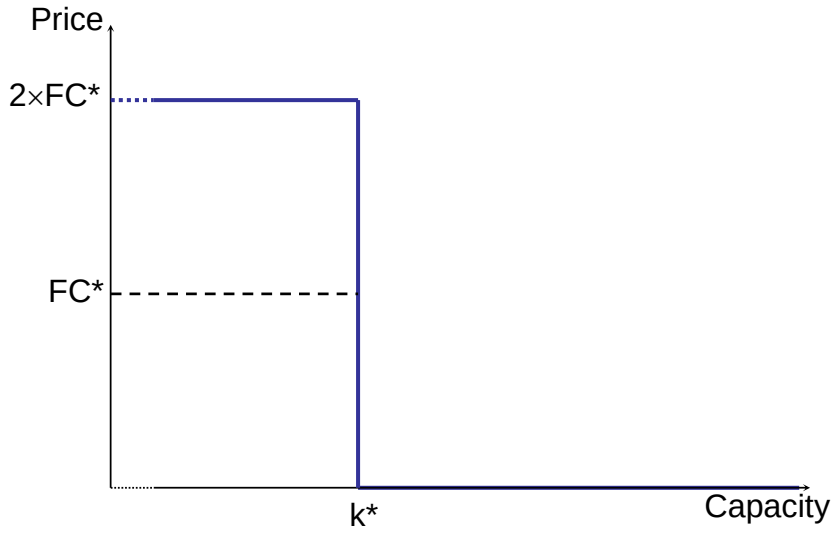
Clearly, the “dog” plant contributes little to reliability, and its efficient ICAP payment should be near 0%, not 100%, of the full payment.

Elements of a successful market design

A capacity market must induce the right amount of investment of the right type in the right locations. Since it is also possible for a well designed capacity market to reduce risk and market power associated with the energy market, it should do this as well. Inducing the right amount of investment, though always controversial due to the large revenues involved, may actually be the simplest of these goals.

Total capacity. Inducing a particular level of investment, k*, is mainly a matter of paying more than the fixed cost of a peaker when there is less capacity than k* and less when there is more capacity. Of course, one must take into account how much more and how much less and revenues from the energy market. It is simplest to start by considering the sum of ICAP revenues and the energy rents of a peaker, which will be termed peak-energy rents (PER). This combined stream is the fixed-cost recovery or short-run profit (SR_π) of a peaker, and should be the initial subject of design. Short-run profit is a function because it varies with the amount of installed capacity. (ISO-NE calls this the ICAP demand curve.) The most obvious shape for $SR_\pi(k)$ is a step function, as in Figure 2, which goes from 2 FC* on the left of k* to zero to the right, where FC* is the long-run annualized fixed-cost of a benchmark peaker. If, over time, capacity, k, is symmetrically distributed about k*, then it will earn, in expectation, FC*.

Figure 2. Inducing a level of capacity



The argument that we should expect this function to induce k* (provided FC* includes the appropriate risk-adjusted return on capital) runs as follows. If investors expected k to average less than k*, then they would expect actual fixed-cost recovery to average more than FC*. In this case they would invest more than they are investing. In equilibrium they will expect k to equal k*, so we should too, since we cannot possibly have more accurate expectations on our own than do investors. The principal drawback to

the use of this step function is that it causes short-run profits to be riskier than necessary, which will be discussed shortly. The cure is to reduce its slope near k*.

Capacity type and behavior. Inducing the right type of capacity is more difficult, but an indirect approach provides a solution. The shortfall of capacity occurs because, when capacity averages k*, peakers cannot cover their fixed costs. If they could, they would invest until k = k*, and there would be no shortage. This proves that the missing energy market revenues are missing from the hours when price rises above the marginal cost of a peaker. In a competitive market, prices are this high only when all peakers are in use. These times will be referred to as the peak hours.

Traditional capacity-market designs essentially pay back the missing peak-hour revenues during all hours of the year. Since during peak hours, the price is above the marginal cost of a peaker (the highest-marginal-cost generator the market would build) it is also above the marginal cost of all but some out-of-date generators that would no longer be built. In fact it is higher than the marginal cost of all but a few old peakers. Hence peak-hour revenues should flow to almost all generators in the market. If this were precisely true, paying peak revenues in proportion to megawatts of capacity installed would be equivalent to paying revenues on peak. This is the assumption behind traditional capacity market designs.

Reality is more complex. Peak-hour energy prices send different signals than per-megawatt capacity prices. If the additional capacity had been induced by a well functioning, unregulated power market, the additional revenues would have flowed through the energy market, and as just argued, those revenues would have flowed during peak hours. Since these prices would have been determined by supply and demand, we can be sure the signals sent by these prices would have been signals for efficient behavior. In practice, such behaviors include, keeping your coal unfrozen in the winter, having appropriate gas contracts in place, expediting parts delivery to recover quickly from a forced outage, and improving maintenance. All of these are actual examples, but what may be more important is the examples we have not yet thought of, but which well-motivated suppliers will think of. Paying them regardless of performance will not provide such motivation.

Besides these short-run examples there are long-run examples. On-peak payments will reward generation that can start quickly and ramp up quickly as well as baseload generation, which is online most of the time. Conversely, older generators, near retirement, will be signaled to retire sooner if they cannot provide capacity when it is most needed and later if they can. This is exactly as it should be. Hence a proper capacity-market design will return the missing revenues to the generators by making capacity payments during peak hours.

Capacity location. The location of new investment is important. The fact that peak-hour energy revenues are missing indicates that the locational signals of nodal energy pricing have been distorted. It might be thought that making capacity payments on-peak would restore the missing part of the nodal-price signal. That would be true if capacity

prices were simply proportional to some energy price-spike that had been reduced by some constant fraction, but the distortion of energy price is more complex.

Because of these difficulties, the market must be divided into zones for which different k*'s can be estimated. In fact this is already common engineering practice. This approach can only approximate the ideal locational signal, but given that some detailed locational signal is still provided by the energy market, this approach should work reasonably well. Once different k*'s have been estimated, these will determine different short-run profit functions and different capacity payments in the different zones. These different payments are determined by treating the short-run profit functions as demand curves (hence ISO-NE's name for them), taking note of the transmission constraints between zones, and computing "zonal prices" as competitive equilibrium prices subject to zonal transmission constraints. These zonal prices serve as the capacity-market's setting of short-run profits.

As an example, suppose the Boston zone has 4 GW of capacity, and its short-run profit function indicates $9/kW-month of revenues when capacity is this short, but there is plenty of capacity in the rest-of-pool zone and the price there would be $2/kW-month. The price difference will cause a maximal "inflow" of capacity over the available transmission and this will result in a lower "zonal price" in Boston, which is read from Boston's demand curve at 4 GW plus the transmission limit.

Risk reduction. Capacity markets are inherently less risky than energy markets because their prices fluctuate only based on the level of installed capacity and not because of weather. Energy-market prices suffer fluctuations for both reasons. Moreover, with proper implementation, capacity markets can hedge the remaining weather-related fluctuations of the energy market. In a hot year with higher-than-normal average PER, capacity market payments can be reduced by the increase in PER. In this way all annual weather-related risk can be removed from the market.

Another feature of energy market revenues is that their expected value is a steep function of capacity in the vicinity of FC* and they do not flatten out on the left like the step function considered above. In fact they increase exponentially on the left. The result is that a symmetrical distribution of k that produces FC* on average will do so by providing several times FC* in a few years and much less than FC* in many years. Investors find such income streams risky. Long-run contracts can eliminate this risk if they cover an investor's total costs. But not all investments can be covered this way, otherwise there would be no meaningful spot-market price to arbitrage long-run contracts against. In practice, investors face considerable exposure both to spot market prices and to long-run contract prices based on predictions of future spot market prices. The result, at present, is an investment market with a high risk premium, especially compared with regulated markets. This is simply a dead-weight loss, and one that can be considerably reduced with a well designed capacity market.

If the short-run profit function is given a moderate slope so that short-run profits decline gradually over, for example, a 15 percent range of installed capacity, profit fluctuations can be considerably dampened. This reduces the remaining investment risk.

Market power in the spot energy market. A beneficial side-effect of subtracting actual PER from the capacity payment is that it eliminates much of the market power in the energy market. Typically this is exercised when capacity is in short supply which means the energy price is already at the marginal cost of a peaker. In this state, market power will inevitably push prices higher and result in an increase in PER. Any generator receiving capacity payments will find this increase deducted from its capacity payment. Consequently most exercises of such market power will no longer be profitable.

Features of the New England design

The New England capacity market will be monthly, and although a good case can be made for an annual market, there are theoretical advantages to a monthly market when neighboring ISO's have different annual load profiles. For example, if one is winter peaking and the other summer peaking, it makes sense to sell capacity into one in the winter and the other in the summer. This is difficult for an annual market to arrange.

The New England market is bordered by the NYISO, which unfortunately does not recognize that the value of capacity fluctuates during the year. Consequently, ISO-NE has had to consider a second-best solution to match this flaw in NYISO's ICAP market (or New York must reform its market). If ISO-NE had priced capacity economically it would have bought capacity from NYISO during summer months even when it had relatively much more capacity then NYISO. This is because NYISO's summer capacity is under-priced, quite likely by a factor of two or more. While this might be beneficial for ISO-NE, it is not socially efficient.

The monthly market described here is stylized but based roughly on the ISO-NE proposal now before FERC. To reduce seams issues with the neighboring New York market, the design does not recognize seasonal fluctuations in the value of capacity.

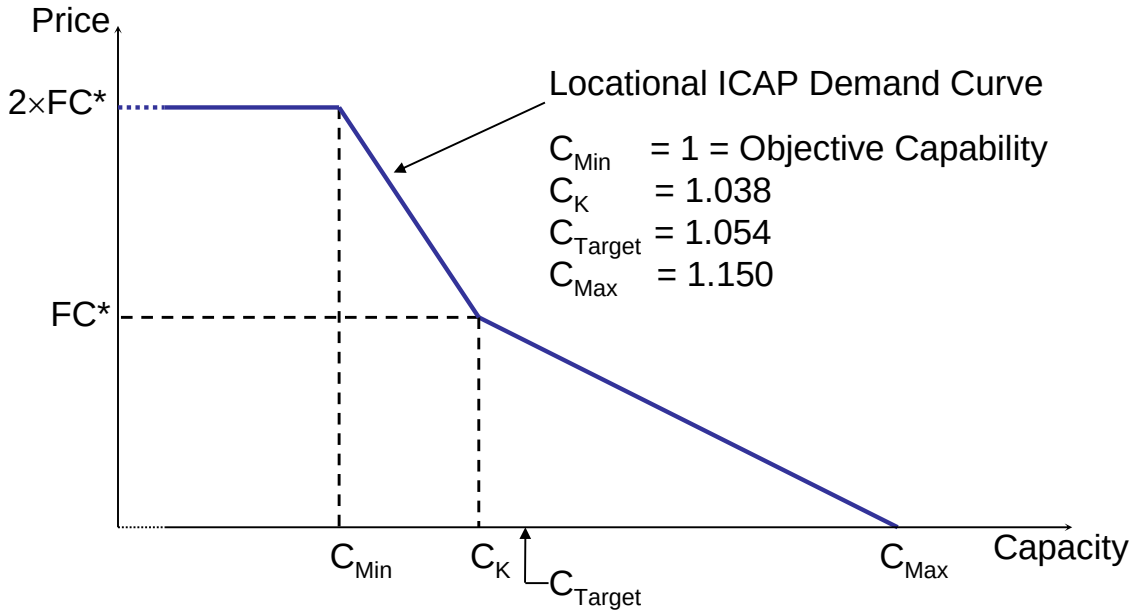
The short-run profit function. The first piece of the design is a short-run profit function. This function should be designed to minimize the total cost of power and un-served load, while inducing some target level of capacity C_{Target}. Hitting the target is relatively easy. Minimizing costs is a problem that requires much data for which there are simply no reasonable engineering or statistical estimates.

The two keys to cost minimization are risk minimization and lost-load minimization. Little is known quantitatively about either of these. We do know that an extremely steep SR profit function is very risky. For instance, one that pays $20,000/MWh below a threshold and nothing above the threshold (rather like VOLL pricing) will produce an erratic and thus risky profit stream. We also know that a nearly flat SR profit function will send weak signals about how much capacity is needed. This will either cause high costs from lost-load if it is too low, or high cost from extreme over building if it is too high.

The curve, shown in Figure 3, which is reminiscent of the short-run profit function in Stoft (2002, p. 178), is quite close to the New England ISO's proposed design. The primary goal when developing this curve was to minimize the chance of serious error. The expected carrying cost of the benchmark peaker (FC*) plays an important role, since

this is the profit that needs to be recovered on average for an investment in capacity to be marginally profitable. A second critical input is the objective capability C_{Min}—the amount of capacity needed to keep loss of load events that result from insufficient resources to a frequency of 1 day in 10 years. All capacity measures are scaled relative to C_{Min}.

Figure 3. "Locational ICAP demand curve" proposed in New England



With FC* and C_{Min} in hand, the demand curve is nearly nailed down from the following three conditions:

1. $SR_{\pi}(C_{Min}) = 2 \times FC^*$.
2. $SR_{\pi}(C_K) = FC^*$
3. $(C_{Max} - C_K) / (C_K - C_{Min}) = 3$

There, however, remains one condition, which will determine C_{Max} (and thus C_K from condition 3): the expected value of SR profits over the distribution of capacities equals FC*, so the benchmark peaker just breaks even in the long run. To compute this expected value, we assume that capacity is normally distributed with standard deviation SD, which is estimated from prior years, and mean $C_{Target} = C_{Min} + .94$ SD. This implies a 17% chance that capacity will be below the amount required to meet the 1-day-in-10-year standard. In New England, $C_{Target} = 1.054$. The expected profit condition, then, implies the capacity price falls to zero at $C_{Max} = 1.150$, and the kink in the demand curve occurs at $C_K = 1.038$, just to the left of C_{Target}.

Subtracting actual peak energy and reserve rents. An interesting aspect of this design is that the curve shown above is not the ICAP payment curve but the SR profit function itself. More specifically, it is the SR profit function for the benchmark generator

which, in this case, is a peaker (a "Frame" gas turbine). SR profit, SR_π, is the sum of peak energy and reserve rents, PER, and the ICAP payment, ICpay. In order to control SR profits according to this function it is necessary to set ICpay equal to SR_π – PER.

The NYISO has been explicitly subtracting something like PER from its equivalent of the $SR_\pi(k)$, since 2003, so the basic idea is not new. But NYISO estimates PER in advance for unspecified but presumably average conditions. With few years of data to work with, and with market rules changing frequently, this estimation process is murky, fraught with difficulty and controversial. The main problem is, of course, predicting energy price spikes which are known to be erratic and unpredictable.

ISO-NE uses an ex post approach. It simply waits until prices happen and then subtracts the PER based upon the actual energy prices. This is much easier and less controversial. Still some difficulty remains because, when the energy price jumps from $50 to $500 for three hours, it is difficult to estimate how much of that price spike would be captured by an actual benchmark unit. It might be on line already, or it might take half an hour to start. Or it might not start for an hour because it spent half an hour trying to decide if the price spike would last more than half an hour. Fortunately the behavior of peakers is much more easily studied and far more predictable than the behavior of price spikes. In any case, this dispatch problem must be solved whether the NYISO or the ISO-NE approach is used.

Besides simplicity and accuracy, ex-post PER estimates have two major economic advantages. First, they reduce energy-market risk, and second they reduce energy-market market power. Consider a typical year with $C = C_K$. In the ISO-NE market, a supplier can be sure that the benchmark generator will earn $SR_\pi = FC^*$ no matter how the summer weather turns out, what nukes go out, or who exercises market power. In an ex-ante PER design, the ICAP payment will be known with this same certainty, but PER will be as uncertain as always. Thus SR_π will be as uncertain as always, which is considerably riskier than having SR_π known in advance.

Besides stabilizing SR profits, ex-post PER reduces market power. Just as real scarcity caused by hot weather cannot raise SR profits, so artificial scarcity caused by withholding cannot raise SR profits. Any increase is just subtracted out of ICpay. Since market power often raises price above the marginal cost of a peaker, much energy-market market power will be eliminated by the ex-post PER calculation.

The reduction in market power would be more complete and the calculation of PER simplified if all prices above the marginal cost of the benchmark peaker were counted towards PER, thus assuming that the benchmark peaker is available for energy whenever the price exceeds its marginal cost. Real peakers are not so reliable, so this would require that a correction factor be applied to the calculated PER, based on several years of past peaker performance. Such an approach is recommended here.

Annual smoothing of PER. PER will fluctuate dramatically during the year, but under the current design SR_π does not. As explained earlier this is to match the flaw in the NYISO design. If monthly PER were used, this would result in ICpay being least in

August when PER is greatest. To avoid this, PER is averaged over the twelve months prior to the monthly ICAP auction. In this way it stays reasonably flat as does SR_π and it is known with certainty at the time of the auction. Actual PER is still subtracted but over a twelve month period.

Relationship to forward reserve market. New England is unusual in that it has a forward reserve market to procure offline reserves well in advance of the spot market. This market has proved essential in motivating the supply of flexible resources during a period without an efficient capacity market. Following the introduction of LICAP, the forward reserve market will play a less important role in providing incentives for flexible resources, since these resources will be substantially rewarded by LICAP. However, the forward reserve market will continue to provide additional compensation to reserve resources to the extent they are undersupplied in a particular location. Most importantly, the forward reserve market sets a locational price for reserves well in advance of the spot market. In this way, the price reflects the economic costs of reserve supply from units other than quick start generators, such as dispatchable load or slow-start units that provide reserves from ramping.

Availability scoring. As explained above, when there is sufficient capacity for reliability purposes, suppliers make too little. This means peakers make too little, which means revenue is missing from peak hours. This is the failure that the ICAP market must remedy, and it should do so by returning the money during peak hours. In theory it should be paid exactly when price distortions occur. The simplest theory would say this happens only when the price hits the cap, but it appears that ISOs engage in more complex price distortions than this. What is certain is that peakers must be missing revenues for fixed cost recovery, otherwise investors would build peakers until there was a reliable level of capacity.

From a practical point of view, one wants to pay the installed-capacity payment, ICpay, only when the incentive for more supply is clearly a beneficial incentive. When the system runs short of NERC-required operating reserves, there can be little question that more supply is called for. Thus a rule that pays more only when the system runs short of operating reserves is a safe rule from an efficient dispatch perspective. Similarly, there would seem little harm to inducing more supply when the price is expected to hit $1000. In fact inducing more supply any time the price is high enough to induce every bit of available supply seems quite safe. Hence, if the most expensive old generator on the system has a marginal cost of $300/MWh, then all hours with higher prices or operating reserve shortages can be declared “shortage hours” and used for paying ICpay.

The next question is how to implement payment during these hours. If there are two shortage hours in a month with an ICpay of $8,000/MW, then each hour can be assigned $4,000/MW and every supplier is then paid an extra $4,000/MWh for each MWh of energy or reserves delivered during these two hours. This has two problems. First, it is risky for the suppliers. Second, most months will have no shortage hours. A less direct approach solves both problems.

Generating units are given an Availability Score, A_G, determined by the fraction of shortage hours during which the generator is available. If this score is 90%, then the generator receives 90% of ICpay. The Availability Score is determined using an exponentially weighted moving average, which is accomplished by proportional updating as follows.

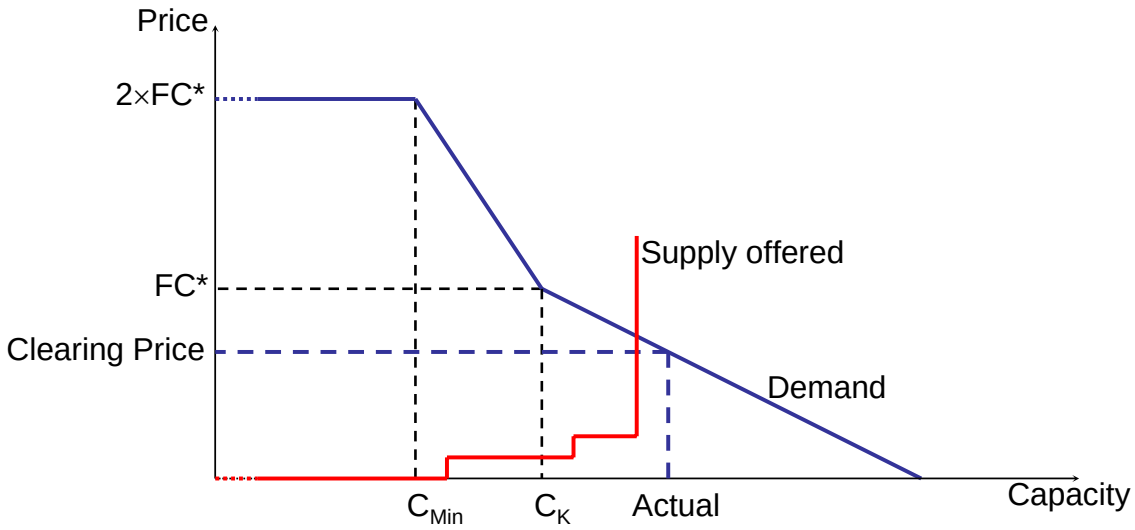
Availability updating. A_G is updated following each shortage hour based on the old A_G and A_{HG}—the actual hourly availability during the shortage:

New A_G = (95 × Old A_G + 5 × A_{HG}) / 100.

This is repeated for every shortage hour since the last update in the order of its occurrence. A_G is initially set at the unit's EFOR'd value. If no shortage hours occurred, then A_G is not updated. All generators are updated according to their performance and without regard to their trading position in the LICAP, or any other, market, except for ICAP importers, which are evaluated only during months in which they sell ICAP to the ISO.

Note that when a generator is available in a shortage hour its A_G is increased by 5% relative to what it would have been had it missed the shortage hour. This increase affects all future hours. Because of the exponential weighting, the total impact is exactly 100%. Because losing future income is less costly than losing present income, the incentive is somewhat weakened, but is strong nonetheless.

Figure 4. LICAP market clearing to eliminate market power



ICAP market power. ICAP markets have steep demand curves and can have significant supplier concentration, especially inside of constrained zones. Hence, market power can be a serious problem if the market is not designed correctly. The trick is to simply ignore supplier bids when computing the market price, as shown in Figure 4, and base the price on the actual capacity in the market. Since supply bids determine who supplies, but not the price paid for supply, no supplier can affect the market price by withholding supply, and hence no supplier has market power. Since supplier bids do not

determine price (SR_π), what does. The answer is ICAP. After all, controlling the level of installed capacity is the sole motivation for paying ICpay, so the appropriate feedback for controlling price is the capacity variable. Fortunately, the ISO knows exactly how much ICAP exists each month, as does the market, so the process is completely transparent.

Note that this is not a market power mitigation mechanism. In such a market, suppliers *have no market power* so mitigation is not necessary. Compared with fixed price caps or elaborately variable price caps such as used in NY and sometimes in CA, this approach is a non-invasive way to deal with potential market power problems.

Supplier bids still perform the function suppliers need them to perform. They can no longer use them to control the market price (exercise market power) but they can still use them, as they should, to signal whether they wish to sell at the market price. They are price takers, exactly as in a competitive market.

Conclusion

Capacity markets are needed in today's restructured electricity markets.[3] This need arises because current power markets have no ability to sell reliability and the high administered shortage prices required to induce a reliable level of capacity are generally suppressed by various market-power mitigation measures. By restoring the missing peak energy revenues, capacity markets attempt to create efficient investment incentives. However, current capacity markets have serious weaknesses. These weaknesses will likely lead to the failure of the markets, if the designs are not fixed.

The New York capacity market is an example of one of the better capacity markets. It uses a downward sloping demand curve to determine the capacity price, and this price is determined on a locational basis, recognizing transmission constraints. Thus, the capacity compensation is responsive to locational supply scarcity. This is a good thing and is the primary reason that the New York market appears to be working well. However, the New York market has four major flaws:

1. Suppliers in constrained zones have an incentive to exercise market power in the LICAP market.

2. Suppliers have an incentive to create real-time shortages.

3. Peak energy rents are estimated without regard to the capacity level or the actual energy prices.

4. Generating units are paid LICAP even if they are unable to supply energy or reserves in shortage hours.

The proposed market in New England addresses each of these flaws:

[3] A requirement for energy options at various strike prices is a viable alternative. See Chao and Wilson (2004) and Oren (2004). This approach, if implemented at the state level, is problematic in electricity markets like New England that include many small states, since a lack of coordination among states likely would introduce market distortions.

1. The capacity price is determined from actual capacity, rather than bid capacity, so no supplier has an incentive to exercise market power in the LICAP market.

2. Ex post peak energy and reserve rents are subtracted from the LICAP price. Thus, a LICAP supplier does not have an incentive to create real-time shortages—the high shortage price resulting from the shortage is subtracted from the LICAP price, so there is no net gain from the high price. This eliminates the second and third flaws.

3. LICAP is paid to suppliers based only on their demonstrated ability to supply energy or reserves in shortage hours. Thus, only supply that contributes to reliability is rewarded.

PJM has proposed a capacity market with much different timing than either New England or New York. In PJM, the capacity auction occurs four years ahead for a one-year capacity product. This long lead time has the advantage that potential new entrants can compete with incumbents in offering capacity. However, such a design means that market power in the capacity auction cannot be addressed as simply as in the New England approach. Moreover, it is unclear whether the four-year-ahead one-year price signal will be a better motivator of efficient investment than the monthly signal in New England, which is easily estimated even several years in advance from estimates of New England capacity.

One might infer that the New England capacity market is a means of allowing a low energy price cap and yet still provide compensation to generators to motivate sufficient investment. This, however, is the wrong interpretation. In our proposal, the generator receives a large reward—much more than the price cap—for providing energy and reserves during shortages. Further, the market hedges the administrative shortage price. This means that little volume is actually transacted at the shortage price, which should enable setting a higher shortage price more in line with the value of lost load.

Two essential elements of good market design are illustrated in the New England approach. First, careful attention to product measurement means that suppliers have the correct incentives to supply what load values: capacity that supplies energy or reserves at times of shortage, and thus capacity that contributes to reliability. Second, market power in both the energy and capacity markets is addressed in a simple and robust way. The end result should be a capacity market that all participants can trust to lead to efficient behavior both in the short run and the long run.

References

Besser, Janet Gail, John G. Farr and Susan F. Tierney (2002), "The Political Economy of Long-Term Generation Adequacy: Why an ICAP Mechanism is Needed as Part of Standard Market Design," *Electricity Journal*, 15:7, 53-62.

Chao, Hung-po and Robert Wison (2004), "Resource Adequacy and Market Power Mitigation via Option Contracts," EPRI, Palo Alto, CA.

Cramton, Peter (1999), “Review of the Reserves and Operable Capability Markets: New England’s Experience in the First Four Months,” White Paper, Market Design Inc.

Hobbs, Benjamin F., Javier Inon, and Steven Stoft (2001), “Installed capacity requirements and price caps: oil on the water, or fuel on the fire?” *Electricity Journal*, 14:6, 23-34.

Oren, Shmuel S. (2004), “Ensuring Generation Adequacy in Competitive Electricity Markets,” EPRI, Palo Alto, CA.

Stoft, Steven (2002), *Power System Economics: Designing Markets for Electricity*, New York: John Wiley and Sons.