Auction Design for Capacity Markets

by

David Salant, Toulouse and Columbia

and

Robert Stoddard CRA International

DRAFT

June 10, 2008

Abstract

This paper reviews auction design issues for new capacity markets. We begin by describing the need for some mechanism to provide the "missing money" needed to ensure desirable level of system reliability. We then consider auction design options for ensuring competitive auctions, including versions of supply scheduled, sealed-bid and clock auction designs. This analysis of auction design options includes an assessment of bidder incentives. Further, we address how auction design can mitigate potential impact of market power, and consider other remedies where market power is known to exist. One of the major issues facing participants in the process is that utilities can want to reserve the option to purchase capacity outside the centralized market. We discuss ways this can be accomplished while preserving, and potentially enhancing, participation and competition in the centralized market. In particular, we explain how this transforms the centralized market from a one-time auction to a set of sequential auctions. We then explain how the price can be expected to vary across auctions and tenders, based on decisions made to shift purchases across channels.

1. Introduction

Regulatory policy has resulted in the introduction of "resource adequacy" or *capacity* markets into the energy sector over the past decade. This paper first examines the need for capacity markets, and second, the types of markets or auction mechanisms that can best serve those needs. This paper will explain why capacity markets are a crucial policy tool to ensure market mechanisms can provide adequate system reliability. The alternative to capacity markets is more direct regulatory intervention. Capacity markets are, as is explained below, a creation of regulatory agencies, and the types of capacity markets being introduced would not exist without a regulatory mandate. Given the genesis of capacity markets, we describe markets that can effectively address those system reliability objectives which are the reason for capacity markets in the first place.

This analysis applies to a particular type of capacity. The generation capacity addressed in this paper is not the physical assets used for the generation of electricity, nor entitlements to the power associated with that generation.[1] Rather, the capacity which is the subject of this paper can probably best be characterized as *administrative capacity credits* that impose an obligation on physical generation to make its power available for sale at market prices but not the obligation to serve load, that is, sell electricity to retail customers, at an agreed price.

[1] For simplicity, we use the term "generation" in the place of "capacity resources." Some power pools have rules that allow capacity credits to be granted to controllable load, inter-regional transmission, and other resources. Our conclusions generalize to all such assets.

As explained in more detail below, the capacity credits are intended to capture an externality: costly capacity resources need to be available to system operators to meet reliability criteria, but each customer has an incentive to under-procure its share of those resources (except in the special case where there is only one customer).[2] Whether the price of capacity credits matches the value of this externality (and over what timeframe) depends on how the capacity credit market, or auction, is organized. The main objective of the is paper is to describe how the structure of the capacity auctions affects the expected value of the equilibrium price, and to compare this expected value of the equilibrium price ("EEP") with the value of the externality. This paper does not provide an explicit analysis of optimal auction design. Rather, we start with a few benchmark proposals and then examine how variations of those proposals will compare to the objective of determining the cost of the *missing money* needed to price the externality.

The market structures we analyze are based on those prevailing in parts of the US. In many parts of the US, a regional transmission operator ("RTO"), such as ISO-New England ("ISO-NE"), the New York Independent System Operator ("NYISO"), the PJM RTO, or the California Independent System Operator ("CAISO") have the mandate or authority to ensure resource adequacy.[3] The means taken to ensure resource adequacy in

[2] In most industries, suppliers are wholly compensated by the sales price of their product. In electricity markets, however, the imposition of side-constraint through reliability rules, combined with various market rules governing the sale of electricity, generally requires a larger capital stock than can be supported by energy sales alone. Hence, in order to stimulate sufficient investment to meet these reliability standards, some implicit or explicit side-payment for "capacity" is needed in this case. See Stoddard 2008 at 13-16.
[3] The State of Connecticut has challenged the jurisdiction of the Federal Energy Regulatory Commission to establish mandatory RTO resource adequacy markets, but regardless of the proper venue for such markets, the externalities related to the positive effect on additional resources for reliability must be addressed through some such market mechanism if, as in the case of each of these RTOs, there is no common regulator mandating compliance of each customer.

many jurisdictions has been the creation of tradable capacity credits, which any entity serving load must acquire.

The market mechanisms we describe are new, and involve a transition from a regulated approach. An essential component of the auction proposals addresses in this paper are the transition provisions.

The next section describes the justification for creating administrative capacity markets.[4] The following section describes the main issues in the transition from a regulated approach for energy supply to a market based approach. The transition issues have two main components: the first is how to compensate old capacity relative to new capacity, and the second is that there is little experience with capacity markets that can be used to guide significant auction design decisions. Auction theory can, as is explained in Section 4, only provide limited insight into very significant decisions that need to be made. Section 4 discusses the relative merits of a few alternative auction designs, in use or being considered. Finally, Section 5 concludes.

2. The *Missing Money*

[4] Purists may argue that these constructs are not 'markets' because there is no willing buyer. Regardless of the name used, these constructs use market-like mechanisms to price and allocate capacity credits and so we refer to them as "markets,"

This section describes the *missing money*. Whether there is a need for capacity markets has been debated in the economics literature and in regulatory proceedings.[5] This section provides two justifications for a capacity market, one based on economic principles and the other, which is consistent with the first, based on practical concerns.

Restructuring and deregulation has resulted in many jurisdictions, in separation, either complete or structural, of the distribution utilities from generation. In most jurisdictions, independent power producers can compete more or less directly with utilities for customers. As a result, restructuring and deregulation have resulted in regulated utilities serving only a fraction of the total load from their own generation, and so energy generation services are being provided by independent and competing entities. When two or more entities supply energy, the responsibility for ensuring reliable energy supplies is necessarily a joint responsibility among all the suppliers since, given current technologies, reliable electric service to *any* end-use customer requires a reliable grid for *all* customers. Should one firm fail to provide adequate supplies in times of peak energy demands, all suppliers customers will suffer from supply interruptions. This is a classic externality.[6]

There are several approaches going back to Pigou and Coase for addressing externalities. Coase argued that what was essential was assigning the "property right" to ensure efficiency of the market. As we explain below, establishing tradable capacity credits is a

[5] See, e.g., Hogan (2005) and Hogan (2006) for a discussion of "energy-only" alternatives and BGS (2007) for a discussion of purely administrative alternatives.
[6] While, in theory, customers of any deficient energy supplier could be selectively curtailed, there are technological and regulatory barriers to doing so in practice.

form of property right to correct for the externality. However, implicit in Coase's Theorem about externalities is that the market for capacity operates competitively. This may not be the case when capacity is lumpy and differentiated by location, ownership is concentrated, large utilities dominant the buy-side, and trading is infrequent. This section assumes that the capacity market operates competitively. Section 4 below discusses how auction design affects efficiency of a capacity market.

To understand the role of capacity market, consider a simple example in which there are two retail suppliers in an area, each serving no more than 100 MWs for an average of 8755 hours per year,[7] but that in a high-demand summer they will each face demand of 110 MW for up to 5 hours per year. If one firm purchases the capacity to serve 110 MW, but the other does not, then for those 5 hours per year, the customers of both firms will suffer service interruptions. Part of the benefit of the last 10 MWs of capacity for each firm accrues to the other firm and its customers. Absent some capacity market construct, however, neither firm will receive compensation for this increment to reliability.

Suppose that some regulator requires that each firm control or acquire110 MWs of capacity in order to ensure reliability even in high-load years. The last 10% of the physical capacity may not run in an average year. Absent payments for standby capacity credits, that capacity will only receive payments for the energy it produces in a few hours every few years. The energy price needed to allow those plants to recover costs would

[7] There are 8760 hours in a 365-day year.

be quite high—probably in excess of $50,000/MWh, even though the marginal production cost of power from these units would only be about $100/MWh.[8]

These episodic price spikes might not be a problem, and can be alleviated by non-invasive means, except for the fact that system reliability is a public good. Absent this externality (and setting aside for now any market power issues), utilities, competitive retailers, or end-use consumers can purchase contracts to protect against price spikes. Since this contract need be only financial, a competitive market can be relied upon to price the "insurance" at efficient levels, for if one firm charges too much for a firm energy price, another can step in and undercut it.Sellers of these contracts, in turn, would rationally choose to hedge their position with physical, deliverable generation which, in turn, ensures a market-based level of reliability.

This logic may fail at several steps, however. Not all consumers may choose to purchase a firm price guarantee for their full expected load. Similarly, sellers of the firm energy contracts may not fully hedge their short position with physical generation. Even with ideal spot pricing of energy, these strategies may be individually rational; the likelihood of under-hedging is increased, however, because all RTOs have price caps well below the level needed to make energy payments alone compensatory for the full number of resources needed to assure reliability. Consequently, it is less costly for financial intermediaries to fail to deliver power and pay the associated penalties rather than to secure sufficient generation. Moreover, since load curtailments cannot readily be localized to particular customers who have not contracted for power, some consumers (or

[8] Hogan (2007).

their retail suppliers) may prefer to 'free ride' to some degree on the capacity supplied by others. Since the capacity purchases by other then provide both private and public benefit, the implicit insurance premiums in the energy contract costs will not cover these external costs, unless some mechanism is in place to ensure reliability contributions from *all* consumers equally (or the ability to curtail preferentially customers who have secured inadequate supply).

A capacity payment provides both a floor on the level of installed capacity and supplemental payments to developers. The floor ensures that installed capacity is sufficient to meet forecast peak demands and engineering analysis of reliability requirements. The payments provide an incentive for developers to invest in generation capacity beyond what would be provided absent the capacity payments. If the market is efficient, then the capacity prices would settle at a value just sufficient to induce investment in capacity up to the target levels. The efficiency of alternative capacity market designs is the subject of section 4.

The other, more practical, reason for capacity markets is that some provisions are needed to ensure that load serving entities have resources available to serve the needs of their customers. Very high prices for a very small fraction of hours are needed to compensate investment in capacity to meet peak load requirements. A pure market approach will not achieve optimal resource adequacy due to externalities described above. However, a regulatory agency, or the RTO, can impose penalties on load serving entities who fail to provide adequate generation. However, the penalties may need to be high, and in any case such fines lead to an implicit market to exchange capacity to avoid the fine, but such ad hoc markets may be much less efficient and effective than intentionally designed ones.[9] So, as a practical matter, it is likely more efficient to use a "carrot" of capacity payments rather than a "stick" of fines and penalties for failing to meet resource adequacy requirements.

One other issue warrants mention. Despite the regulatory restructuring that has been ongoing, in most states, and in many other countries, default service, or standard offer service remains a regulated service. The utilities responsible for purchasing the energy to meet the load requirements are usually subject to regulatory prudence review, to ensure that rates are based on competitive costs. Moreover, in most cases, the utilities are under a regulatory mandate to purchase energy and ancillary services by means of some sort of competitive bidding process. The costs of the capacity to ensure reliability are a component of energy services that can be subject to regulatory prudence review.

3. The Transition from Regulation to Markets for Ensuring Resource Adequacy

In transition from regulated resource planning environment to market mechanisms for ensuring resource adequacy, decisions are needed about how existing capacity owners will be allowed to benefit from capacity markets. On one hand, the existing capacity is already in place and has incurred fixed (now sunk) costs. So no supplemental capacity

[9] This was in fact the experience in PJM and ISO New England; FERC determined in both cases that the older structures were not just and not reasonable.

prices are needed to provide incentives for construction of these resources, although it is still possible that, even if the going-forward costs for existing units are *lower* than the all-in cost of a new replacement unit, those going-forward costs are higher than the expected net margins in the energy markets.[10] From a rate-payer perspective, this would argue in favor of excluding the existing capacity from receiving any capacity payments or setting up one market for new resources and another for existing resources.[11] Indeed, some utilities have argued in favor of discriminatory pricing approaches, paying existing capacity less than new capacity. Utilities' concerns about high capacity prices is understandable; total capacity payments in the mid-Atlantic area of the U.S. for the twelve months beginning June 2008, paid to nearly 183,000 MW of capacity, will exceed $16 billion.

One approach is to conduct negotiations or auctions for existing capacity separately from auctions for new capacity. The logic underlying this approach is the view that if existing capacity owners have no other recourse then they will presumably accept a lower capacity price than would new developers. For this reason, utilities have, at times, e.g., in California, argued in favor of separate negotiations for the existing capacity; the success of such negotiations in achieving lower capacity prices depends the ability of the buyer to effectuate price discrimination either through monopsony power, lack of price transparency, or the ability to overbuild the market relying on recovery of these out-of-

[10] Since we regularly observe, even in fully regulated systems, that generation units are retired and replaced by new resources, we can infer that the going-forward costs of these retiring units, net of their net energy revenues, are *higher* than the all-in costs of the replacement units, net of their net energy revenues. There is no basis to believe, therefore, that the capacity payment required by the *marginal* incumbent differs much from the capacity payment required by the most economic new resource. Clearly, however, there are many inframarginal existing resources whose net energy revenues more than cover their going-forward costs and, consequently, may be earning positive economic profits.
[11] See, e.g. PG&E (2007).

market costs through regulated rates. One other approach is to require the existing capacity owners to act as price takers, possibly for a fraction of the auction price,[12] or allow the new capacity owners to secure longer term price commitments than existing capacity owners can obtain. Separating negotiations or auctions over time for different types of capacity does not necessarily result in lower prices, and can result in higher prices.

Separating the contracting over time for new capacity and existing capacity creates uncertainty for all market participants that lead to inefficient purchase decisions and higher total costs in the long run. By splitting the purchase of a single undifferentiated product, administrative capacity credits, into two distinct procurements, a buyer in the first procurement can underpay or overpay relative to the second. More specifically, suppose a decision is made to purchase some new capacity based on the mistaken belief that the cost of keeping old capacity on line is high. Alternatively, an early decision to keep old capacity in service or to retire old capacity can be made based on an over-estimate, or under-estimate, of the missing money needed to finance new facilities. This is, in part, an auction design issue, which is discussed in more detail in Section 4 below.

One other major concern in pricing old versus new capacity, and one which is most significant during transition from a regulated resource planning process to a market-based process, is that new capacity investment decisions are based on long term capacity price projections whereas decisions about retirements, mothballing, maintenance and upgrades

[12] Where this has been adopted, e.g. ISO New England, existing generation that can demonstrate that its actual going-forward costs exceed this cap are allowed to bid that cost, in effect borrowing a concept from the older cost-of-service regime.

(which have shorter economic lifetimes) are based on shorter horizons. This means that the duration of a capacity pricing (implicit or explicit) contract for new facilities could be much longer than that needed for existing facilities, price otherwise equal. When there is a long history of capacity prices and an attendant confidence in the regulatory stability of the capacity construct, then investment decisions face less risk than during transition; moreover, decisions about facilities constructed post-transition should reasonably anticipate the likely future capacity payments. In the transition phase, which can be many years, the pre-transition capacity will arguably require compensation for whatever asset impairments may have occurred due to restructuring.

One problem with differential pricing for existing and new capacity is the fact that this can distort plant retirement and/or other maintenance decisions. The potential mismatch in prices between the two procurements not only has cost ramifications for purchasers, but it also indicates an inefficient allocation of capital and, consequently, inefficiency in the market. If the capacity payment to existing resources is lower than the payment to new resources, resources will retire prematurely, requiring greater capital expenditures to maintain resource adequacy.[13] Conversely, if capacity payments to existing resources are higher than the payment needed by new capacity, some older generation should have been retired that was not. If all capacity credits are purchased simultaneously through a single market, new capacity resource can compete directly with existing resources,

[13] Some California stakeholders have argued that early retirement of old resources is *per se* a social benefit. If the costs of these older plants or the value of newer replacement plants are not fully valued in market prices, this argument might be correct, but the economically rationale response is to modify market pricing for outputs (emissions, renewable power, dispatch flexibility, etc.) rather than to skew pricing in a different market.

resulting in an efficient and competitive determination of when resources have reached the end of their economic usefulness.

A further advantage of a comprehensive, single procurement for capacity credits is the ability of customers to select a lower standard of reliability rather than pay the market price for "standard" reliability. The resource adequacy targets are designed to achieve some generally accepted level of involuntary service interruptions, typically one day in ten years. Some customers, however, may be willing to curtail energy consumption more frequently if, by so doing, they can pay a lower price for their delivered power (including capacity costs). A capacity auction creates a direct means by which customers or, more typically, demand response aggregators, offer verifiable peak-load curtailment as an alternative to buying more physical generation resources. An efficient auction design can then choose the optimal mix of new generation, existing generation, and demand-side resources so that all *other* customers receive the default level of reliable delivery.

When there are many different kinds of capacity resources offering to sell capacity credits, the question arises as to whether the product is, in fact, uniform, or whether there should be either differential prices for different kinds of capacity resources. The practice to date has combined these approaches. All U.S. capacity markets are locational, recognizing the fact that resources outside key transmission constraints are imperfect substitutes for capacity within those constraints. For example, the capacity price inside New York City is more than twice the price elsewhere in New York State, reflecting the need to maintain a minimum level of capacity inside the city's limiting transmission system. Similarly, PJM originally proposed to pay a market-determined premium to units with particular operating characteristics needed to support real-time system operations.[14]

Other difference in unit characteristics or performance have not to date been priced as separate attributes, but instead captured either by reducing the quantity of capacity credits that a particular unit is awarded (to reflect, for example, the discounted value of a wind generator, which cannot be counted on to be available during peak periods) or through performance penalties subtracted from capacity payments (used in New England, for example, to penalize units that were not available during system emergencies). The advantage of using these administrative adjustments is that it reduces the number of distinct products sold in the market, therefore increasing liquidity and competitiveness as well as making the auction structure more tractable and transparent.

4. Auction Alternatives

There are three main alternative procurement options that have been used in energy sector procurement: multi-attribute negotiations, price only sealed-bids, and open, clock-type auctions. We briefly describe each, and then provide a comparison of these alternatives.

First, we describe those elements and design features that any procurement process must include beyond those of the form of the bidding. .

[14] This premium was subsequently replaced by a commitment to enhance the ancillary services markets to achieve a similar effect.

a. Definition and size of the lots.

The required capacity can be divided into numerous ways. First, there is the size of the lots. The quantities offered by bidders can be equal to the size of the plants, or can be integer MWs of capacity. Lot sizes that do not equal plant capacities can, in theory, create an *exposure problem* for the bidders. More specifically, a bidder with a plant of size K, could offer a price of p, but may only be selected to supply K' < K MWs of capacity at the offered price. In this case, this bidder may not wish to proceed with construction or supply K MWs of capacity.

On the other hand, allowing bidders to submit large, unit-contingent package bids, can put smaller suppliers at a disadvantage. A set of small supplier may be the low cost suppliers for K units of capacity, but not of K" > K. In this case, a supplier who can serve the larger quantity may win despite the fact that the average cost of small suppliers may be less. This can be due to what is called the *threshold* problem.

b. Heterogeneity of the lots.

Another procurement design decision is whether there should be uniform terms, or whether there should be different products. In most cases, capacity is divided into different zones, and all capacity within a zone is treated as producing uniform, locational capacity credits. Capacity of the different types of plants, peaking, base-load, intermediate, are also treated the same. Likewise, no implemented capacity market distinguishes between renewable generators and other generation, or between flexibly operable plants and others. Dividing capacity into more types allows the auctioneer to control the mix of capacity, but at a cost of auction complexity and competition.[15]

The heterogeneity of resource needs can require different prices for different types of resources. A capacity market that results in over investment in one type of capacity, e.g., base-load, and under investment in others, e.g. peaking, could fail to meet reliability objectives, and also result in excessive costs. How unregulated market forces affect incentives for investing in different types of resources is not the subject of this paper.

We explain in some detail below some of the additional auction design issues that arise when there is a need to differentiate across time and product types.

c. Timing

The timing of the procurement can matter a great deal. Any new, retirement or expansion capacity decision requires significant lead times. However, different types of investment require different lead times. Moreover, aside from differences in the required lead times, the expected procurement costs can be affected by how the quantity needed is divided over time.

One theorem from the auction theory literature, "Weber's Martingale Theorem," suggests that under some circumstances, the expected price does not depend much on the division

[15] Such control also presupposes that the auctioneer has better information than the supply developers as to the optimal mix of resources on the system. If all products of capacity resources are correctly priced, however, the resulting resources should approach an optimal mix. See Cramton and Stoft (2008).

of the quantity over time, provided there are adequate lead times for all the auctions or requests for offers (RFOs).

More specifically, Weber's Martingale Theorem assumes that there are some number N items to be auctioned (procured) in a sequence of 1st or 2nd price auctions, and that bidder valuations (costs) are identically and independently distributed ("iid"), and that each bidder can win one lot. Then, if each bidder's offer in any round is an increasing function of its type (costs), and prices are public announced after each round, then the prices form a M*artingale*, i.e., $E\{p_m | p_1, p_2, \ldots, p_{m-1}] = p_{m-1}$. (See Milgrom (2004)).

These assumptions are restrictive. What follows is a brief discussion of the significance of these assumptions. The assumption that each bidder can win only one lot can be relaxed to cover the situation in which each bidder, b, may have some number of k_b lots or units to offer available and that the costs are iid. Even when costs are not iid, the opportunity costs of each unit, as compared to acquisition costs, or potential market value, could be iid.

The division of the auction volume across lots can, but need not, affect prices. To see this, suppose that the capacity can be divided among two auctions. Participation and bidding strategy across the two auctions can, but need not, depend on how the capacity is divided among the two auctions. Weber's Martingale Theorem implies that, if the costs of the bidders participating on both are the same, then the expected prices in both auctions should be the same.

However, this Martingale Theorem will not apply if the low cost bidders in one auction are not always the low cost bidders in the second auction. To be more precise, suppose that in the second auction the expected price is the k+1st lowest cost assuming that there are k blocks available in that auction. Suppose costs across the two auctions are the same for all bidders, and that there are j blocks in the first, k in the second. Then the price in the second auction will be the k+1st lowest cost among the bidders who remain after the first auction. If the winners in the first auction won't have the capacity to enter in the second auction, then the k lowest cost bidders will win the first auction, and at a price equal to the (j+k+1)st lowest cost among all the bidders. In this case the Weber Martingale result applies, and expected price is the same across the two auctions, independently how the targeted capacity is split.

Now, suppose that for each auction, the cost of the bidders have some random component, which are the not correlated across auctions, but distributed in the same manner in each auction. In this case the Martingale Theorem does *not* apply. The price in the 2nd auction will be the expected value of the (k+1)st lowest bidder among B – j bidders remaining, where B is the number of bidders, and it is assumed that each bidder has one block or unit available. This would then be the same expected price in both the 1st and 2nd auctions. In a single auction, the expected price is the cost of the (j+k+1)st lowest bidder. These costs need not be the same.

As an example, suppose cost are drawn from uniform distribution over the interval [a,b], j+k = 4, and B = 8. Then in a single auction, price would be a + 6/9 Δ where Δ = b –a. If the auction is split into a first one for 3 blocks, and the 2nd one for one block, then the price would a +2/5 Δ in both auctions. If three blocks are sold in the 2nd auction and one block in the 1st auction, then prices will be a + ½ Δ. Thus, the division of the target volume across auctions can matter a great deal.

d. Size of the blocks

In most energy procurements, bidders must identify specific facilities in advance. Bids are often, but not always, based on the size of the project or plant. When the size of the marginal project can be too small or too large, it may be impossible to exactly meet capacity targets, or capacity beyond the targeted levels will be needed to ensure reliability.

An alternative is to fix the minimum block size, e.g., 25, 50 or 100 MWs. When bidders have some flexibility in project size, or when many have a portfolio of assets, the developer costs of meeting capacity targets can be lower. Provided that bidders are required to have qualified projects or plants backing their bids, tying the bid quantities to the size of specific projects is unnecessary, and can adversely affect costs.

The minimum block size can have other effects. Some auctions have 100 MW minimum, whereas others have 25 MW, or even lower in the case of renewable resources. The larger sizes create entry barriers for smaller bidders. In addition, a number of small bidders may not be able coordinate effectively to compete effective against a larger, potentially higher cost supplier.[16] This threshold problem means that larger block sizes can create bias the outcome against small bidders.

On the other hand, allowing small blocks means that the developer are a larger project can obtain capacity payments for part, but not all, of its capacity. This argues in favor of larger blocks.[17] Various type of package bidding designs have been developed to address this type of exposure problem. The UK's OFCOM has just recently conducted a few such auctions, including one in which the winning bidder was able to submit a package bid for all the blocks in the auction in competition with other bids for individual blocks and smaller packages.[18] The New England and PJM capacity auctions also allow new units to make all-or-nothing offers that are accepted only if the total cost is lowered by so doing - which may result in the auction clearing at a price higher than the offer price of the new entrant.[19]

e. *Off-ramps and timing.*

A load serving entity with a certain need for capacity may not want to risk having to pay a significantly above-market price because of some unanticipated market failure, lack of participation, very transitory price spikes, and other random events. Having a fixed date for the LSE to acquire all its capacity at one time creates some risks. For this reason,

[16]. See McAfee and McMillan (1996) or Loxley and Salant (2004) for a discussion.
[17] See Loxley and Salant (2004).
[18] See www.ofcom.org.uk.
[19] If, for example, at the margin the auctioneer needs 10 MW of additional capacity to meet a total requirement of 1000 MW, and has two competing offers, one for 10 MW at $80 and another for 100 MW at $79, the total first-year cost of accepting the smaller, higher cost offer ($80,000) is less than the cost of paying for the surplus ($86,110). Whether such a rule is optimal in the long-run is unknowable.

what some call *off-ramps* can be desired. Above, we have explained why, absent uncertainty, a fixed division of the quantity across auction can, but need not, affect the expected price. Here we expand on this point to address strategic decisions by the buyer/auction originator to shift quantity across auctions.

There are two reasons why it can be advantageous for the auction manager to leave open this possibility. First, market power *within any individual* auction is determined by concentration of supply among the qualified suppliers. This market power can have little relation to concentration in the local energy markets. Even in a concentrated market, an auction can be quite competitive, and conversely.

Suppose, for example, there is a probability α that any auction, including an initial one, will have high concentration, and could result in an above-market price. If the auction were for the entire fixed amount of capacity needed, then suppose in this situation, we would have a price q .> p where p is the usual market price.

Now, suppose that the auction manager can choose, based on bids during the auction, defer a fraction, f, of the needed capacity to a second auction, one in which the expected price would be $r = \alpha q + (1- \alpha)p$. This price, r, is then essentially a reserve price in the first auction. The reserve price could depend on the quantity deferred in the first auction. What this does is effectively alter the residual demand curve facing the large supplier, increasing its elasticity at least over some range.

This type of adjustment process would have two effects: one is to reduce the ability of a firm with market power in the first auction to obtain a price $q > p$, and second, given that firm an incentive to accept a lower price. Indeed, if the fraction f is large enough, even a monopolist might accept the price of r in the first auction.

Second, if an initial auction occurs during a period of unusually high costs, then the LSE might want to have the option of deferring its capacity purchases in the hopes that costs will return towards more typical levels. If markets are competitive such action might not reduce expected costs and may only delay the adjustment to the increased costs. However, if the procurement is not perfectly competitive, and there are only a few suppliers delaying bidding until the resolution of *ex ante* cost uncertainty can affect *ex ante* expected procurement costs.

Suppose, for example, there are two bidders, who each can serve two units of load, and two units are needed in aggregate. Suppose too, that costs can be high c^h or low $c^l < c^h$. Let $c = \alpha c^l + (1-\alpha)c^h$, where α is the probability of low costs. We suppose that each bidder's costs are independent. So, there are four states after bidders learn costs, (H,H), (L,L), (H,L), and (L,H).

We suppose each bidder submits an offer (price) for one or both blocks, and that the lowest cost combination wins. Ex ante, the price is c. However, ex post, with probability $(1-\alpha)^2$, at least one bidder will have high costs, in which case, the price will be c^h. So, the expected value of the price will be $c^h(1-\alpha)^2 + \alpha(2-\alpha)\, c^l > c$. With more bidders, the inequality can be reversed, depending on relative numbers of blocks and bidders. Thus, when a LSE waits until uncertainty is resolved, costs can end up being lower or higher costs depending on the level of competition.

f. Market Power and Volume Adjustments

When one bidder controls a sufficiently large fraction of the supply in an auction, it has the incentive to withhold some fraction of the supply so as to obtain a higher price for the remainder of its supply. Such strategic withholding possibilities are well documented. For this reason, energy procurement auctions have often imposed caps on the fraction of the supply that a single bidder can offer in any single auction; such a cap in capacity markets is not available, however, because the market is typically pricing nearly all of the generation in a region, and the ownership of that generation is what it is. The challenge is how to design an auction structure that can yield competitive results, regardless of the underlying ownership concentration, with the least reliance on administrative offer caps.[20]

We have previously described how the expected price in a sequence of auctions can, but need not be a Martingale. In other words, expected value of price can be the same across auctions. In addition, we have explained why this result does not hold when bidder costs are not perfectly correlated across auctions. In what follows, we explain how this theorem does *not* apply when a bidder's choice of how much to offer for one lot in a first

[20] We assume for discussion purposes that offers are not subject to oversight by market monitors or regulators and that suppliers do not face civil or criminal penalties for market price manipulation. Naturally, the presence of such controls would substantially modify both the incentive and ability to offer capacity in the ways described.

auction affects the number lots that the auction manager will accept and therefore the price that bidder will receive for other, withheld, lots in a second auction.[21] In particular, we explain how the ability of a ability of even a large bidder to withhold supply can be mitigated when the auction manager can elect to alter the supply in the auction.

First, we describe a large bidder's incentives t withhold supply. Consider a bidder offering a supply schedule, S(q) = C'(q), into an auction. To simplify the exposition, we suppose that the auction is a uniform price one, and let p^0 denote the price and q^0 the quantity, with $p^0 = S(q^0)$, that this bidder could expect to sell should it report its true costs. Now consider an optimal deviation from sincere bidding, S*(q), C*(q) be the (variable) cost function consistent with S*(q) and let p* = S(q*) denote the price that the quantity the bidder would sell at a price p* when it bids S* instead of S. Note, that pq-C(q) ≤ p*q* - C*(q*) for all q that the bidder could expect to sell at a price p when reporting any other supply schedule, S(q). If this condition holds with a strict inequality, then it must be the case that C*'(q) > C'(q) for some such q, as in uniform price auctions, a bidder's supply schedule only influences price if it has the marginal supply at the clearing price. Thus, as bidder can have incentives to over-report costs.

Suppose, for example, there is one dominant bidder, the auctioneer has a fixed, inelastic demand, D, up to some reservation price, R, and there is a competitive fringe of other suppliers/bidders with a supply schedule s(q). Then the dominant bidder will behave as a monopolist facing a residual demand DR(q) = R – s(q), and will submit offers S(q) so as

[21] This discussion in loosely based on work by McAdams (2007), LiCalzi and Pavan (2005) and Allaz and Vila (1993).

to achieve the monopoly outcome. Effectively, the dominant firm can set choose q so as to maximize q(R – s(q)) – C(q).

Using this framework, McAdams (2007), has shown that the auction manager can achieve the efficient 1st best outcome when there is variable demand in a single auction. The reason is that the auction manager can effectively adjust the slope of the residual demand facing the large bidder so that the large bidder will elect to choose the competitive outcome. This may require knowledge of bidder valuations in advance of the auction and/or the ability of the auction manager to respond during the auction to information reported during the auction.

LiCalzi and Pavan have shown that even when the auction manager must commit in advance of the auction to an effective residual demand, the price will be lower, quantity higher, and welfare improved as compared to the case in which the auction manager has to commit to a specific auction volume.[22]

The above does *not* imply that, by delaying purchases, the auction manager can affect price in the first of a sequence of two auctions. The reason being is that offers in the first auction will be affected by expectations about prices in the second, and that losing bids in the first can partially or fully roll-over to the second. A complete model of the strategic implications of the auction manager adjusting auction volume across a sequence of

[22] The LiCalzi and Pavan and McAdams papers formally were models of forward auctions, but the same results apply to reverse auctions.

auctions would include some assumptions about how the outcome of the each auction affects participation in the next.

One case is relatively straightforward to analyze. That is the case in there is a random component to participation, and that in each auction there will be a small fraction of low cost suppliers. To keep things very simple, suppose that with a small enough auction volume in the second auction, the auction manager can ensure that price is close to some competitive benchmark level. In this case, the auction manager facing a dominant supplier has some ability to withhold volume in the first auction and purchase the withheld volume in the second from competitive suppliers. This reduces market power of a dominant supplier in the first auction.

g. Auction Format

This section discusses auction design alternatives. Capacity procurement entails provision of multiple blocks and in different zones. This means than any auction design needs to consider the potential substitutes and complements on both the supply and demand sides of the market. The auction design problem to optimally address these issues in full generality is complex.

In what follows we review alternative designs that have been developed for the types of issues present in capacity procurement for different types of capacity in different zones. Our review examines a few key issue. This approach is motivated in an attempt to identify the main ways in which an auction can *fail* and to suggest provisions to reduce the likelihood of such occurrences.

For example, one key attributes of an auction design is how effectively does it permit bidders to arbitrage price differentials across different units in different locations. Absent such provisions due to restrictions on activity rules or in nature of a sealed-bid requirement, price reversals can, and often do, occur – i.e., more valuable product sells at a significant discount relative to the least valuable ones.[23]

i. Review equivalence

One basic question is to what extent does the auction design matter. A fundamental result of auction theory, the Revenue Equivalence Theorem ("RET"), suggests that in may cases it does not matter much. One version of the RET states that when each seller has only one plant to bid, the costs of the plants are drawn from a common distribution over an interval [c,C], then any auction in which the seller with costs C receives zero expected surplus yields the same outcome, i.e., expected payments[24]. In this sense, auction design does not matter unless bidders can exercise market power, or unless an auction can result in misallocations, such as would be the case when high cost sellers can expect a positive surplus. In multi-unit and multi-object auctions, each seller who has more than one unit for sale will want to consider the impact of its offer for one unit on the price it can expect to receive for its other units.

[23] See Milgrom (2004) or Salant (2004) for examples.
[24] See Klemperer (2004).

ii. SDCA vs sealed bid auctions

The above means that in comparing a sealed-bid, first-price auction with an open descending price auction, there should, in many situations, be no differences in the expected outcome. However, in a sealed-bid, first price auction, a bidder can over or underestimate rival costs, and so bidder miscalculation can be more likely to affect the outcome in sealed bid auction than in an open auction. In addition, when costs of different bidders are correlated, an open auction can mitigate the impact of the winner's curse. These factors mean that the simultaneous descending clock auction (SDCA)[25] has some advantages relative to a sealed-bid mechanism.[26]

The SDCA has some disadvantages as well. First, there is the possibility of strategic withholding. When all bids are disclosed during the course of a SDCA, any one bidder can tell how much of an activity reduction is needed to end the auction.[27] This can provide a bidder with a strong incentive to reduce supply slightly to end an auction. Information disclosure can be limited so as to make it more difficult for a bidder to determine whether its reduction will end the auction. This does not totally eliminate incentives to withhold supply. Volume adjustments, as described above can also mitigate withholding incentives.

[25] See Loxley and Salant (2004).

[26] What follows is a brief description of the SDCA auction format. An SDCA is a variant of the SMR auction – both are multiple lot auctions. The lots can, but need not, be identical. Bidding occurs in a sequence of rounds. Bidders whose bids were topped in one round can improve their offer in the next round, and/or switch to other lots. Bidding continues, and prices continue to increase in an SMR auction or decrease in and SDCA until no one bidder is willing to improve its offer on any lot. Activity rules require bidders to continually improve their offers. In addition, rules govern the rate at which prices increase/decrease between rounds. Note, that in some simultaneous auctions, bidders name prices, whereas in others, such as the SDCA, the auction manager names prices and bidders decide which lots and how many to bid. See Loxley and Salant (2004) for a more complete description.

[27] See Loxley and Salant (2004) or Cramton (2004) for discussion.

A further disadvantage of SDCA's and other simultaneous auctions is the time required to complete the auction. Simultaneous auctions can be completed in under an hour with very careful planning, but can take weeks or months to complete. This is a significant disadvantage in energy procurement, but less so in capacity procurement, as in the former, prices tend to be much more volatile than in the latter. Only recently, have any sealed bid designs been available to facilitate arbitrage by bidders, so as to reduce chances of misallocation.[28]

More specifically, if a bidder has to choose between two projects, and cannot develop both, it might need to guess which might warrant a higher price in a sealed-bid auction when it submits its offer. In contrast, in an open, multi-round auction, the bidder can choose which capacity to offer based on price differentials in the auction. Milgrom's *assignment auction*[29] effectively allows bidders to make contingent bids.

Another potential drawback of SDCAs for capacity procurement is that the calculations are needed to determine overall adequacy of the supply in different zones based on bids in any round. There are complementarities across plants. So, a low cost offer at one location may or may not be acceptable depending on resource availability in other locations. This does not mean that an SDCA is not feasible, only that the complementarities may require backing up to early round bids for some resources when other resources drop out of the bidding.

[28] See Milgrom (2008) and Mueller (1993).
[29] Milgrom (2008)

iii. Supply function auctions

Supply function auctions is another type of auction format has been commonly used for energy resource procurement. In a supply function auction, bidders specify offers in the form of supply functions, often piecewise linear functions or step-functions. The auction manager then adds the offers and calculates market clearing prices. Supply function auctions have the superficial advantage that offers are based on physical units. However, this is only a superficial advantage in that a clock auction can allow block bids sufficiently large or final allocation rules to accommodate large units.

Supply function auction seem to provide suppliers additional market power. Green and Newbery (1992) observed in discussing supply function auctions in the UK that ". . . generators . . . could earn extremely large profits while creating large deadweight losses in a market based on price competition that was intended to keep prices close to marginal costs." The earliest work on supply function games (Klemperer and Meyer (1989) identify the possibility of multiple equilibria, and indeed with imperfect information equilibrium price can remain bounded away from costs.[30]

A simple example comparing bidding strategy of a large supplier in a clock auction and a supply function auction illustrates a main drawback of the latter. Suppose there is one large bidder who would want to offer 80 or 100 MW of capacity, and has a cost of 4 per unit. Also, suppose that there is a competitive fringe, so that there is a probability ρ, that the large firm can sell its entire capacity at a price of 10, and with probability (1- ρ) it

[30] See Larson and Salant (2003).

could sell 60 MWs at a price of 10. Further suppose that this firm can always sell its entire capacity at a price of 8. If this firm can name one price for the first 80 MW and another for the second, it could obtain payoff of 10x60 + 8x40 – 4x100 = 520. If it has to name one price, then the firm will set price at 8, if $\rho < 1/6$. So, in this case, the firm will have an opportunity and incentive to set a higher price for the first 80 MW when it can do so.

Note, that in this example, the firm in a clock auction can choose to drop its offer from 100 to 80 when the price falls below 10. This illustrates another point of comparison between clock auctions and sealed-bid auctions. A sealed-bid auction can be structured to be *strategically equivalent* to a clock auction. This is true even if bidders can benefit from learning about other bidders costs, or valuations, during the auction. For example, suppose a bidder faces cost uncertainty, but believes if at least 50% of other bidders remain in the auction below a price of x, then its costs are likely to be c < C. Otherwise, suppose the bidder believes its costs will be C.

In this situation, the bidder could, in theory, specify a sealed bid containing a contingency based on what fraction of other bidders are willing to match a price of x. If such contingent bidding is not allowed, and if information that reduces uncertainty about is revealed during the auction, then a sealed-bid auction will no longer be strategically equivalent to a descending price clock auction. Allowing contingent bids introduces a lot of complexity into an auction, and may not always or often be feasible.

5. Conclusion

An unanticipated outcome of the move to competitive wholesale electricity markets has been the need for some "make-whole" or "capacity" payment to ensure that enough electric generation is available to meet peak loads. By creating administrative "capacity credits" that can be sold by owners of resources to load-serving entities, combined with a mandatory showing of credits, simultaneously ensures resource adequacy and that all consumers bear a proportionate cost of maintain the common good of system reliability. The market structure in which these capacity credits are traded can have tremendous effects on the resulting path of prices and the efficiency and effectiveness of the capacity credit program in sustaining competitive electric markets.

Modern auction theory and game theory provide insights into ways to design effective capacity markets that yield competitive market outcomes even when conditions are less than fully competitive, such as when the underlying structure of capacity ownership is concentrated or information is incomplete. Many features of capacity auctions, such as the clearing mechanism, timing, creation of off-ramps, volume adjustments to diffuse market power, and contingent bidding, can be improved by careful analysis using recently developed auction theory. When informed by sound economic theory, capacity market structures should be robust and competitive, enhancing the trust placed in them by regulators and market participants. This in turn creates an important benefit to consumers: by reducing regulatory and market uncertainty, the cost of managing risks of new electric generation development decline, as will the delivered cost of power.

References

Allaz, Blaise and Jean-Luc Vila (1993) "Cournot Competition, Forward Markets and Efficiency," *Journal of Economic Theory*, Vol 59, Issue 1, pp 1-16. (pdf)

Cramton, Peter, "Collusive Bidding: Lessons from the FCC Spectrum Auctions," (with Jesse Schwartz) *Journal of Regulatory Economics*, 17, 229-252, May 2000

Cramton, Peter, "Collusive Bidding in the FCC Spectrum Auctions," (with Jesse Schwartz) *Contributions to Economic Analysis & Policy*, 1:1, www.bepress.com/bejeap/contributions/vol1/iss1/art11, 2002

Cramton, Peter and Steven Stoft (2008), "Forward reliability markets: Less risk, less market power, more efficiency," in *Utilities Policy, forthcoming.*

Hogan, William W., (2005) "On Energy Only Electricity Market Design for Resource Adequacy," working paper.

Hogan, William W, (2006) "Resource Adequacy Mandates and Scarcity Pricing, Belts and Suspenders," working paper.

Klemperer, Paul (2004), *Auctions: Theory and Practice*, Princeton University Press

Larson N. and D. Salant (2003), "Equilibrium in Wholesale Electricity Markets", Working Paper Series, University College London.

LiCalzi, M., and A. Pavan (2005): "Tilting the Supply Schedule to Enhance Competition in Uniform-Price Auctions," *European Economic Review*, 49(1), 227–250.

Loxley, Colin and David Salant, "Default Service Auctions," *Journal of Regulatory Economics,* Vol. 27, No. 3, 2004.

McAdams, David, "Adjustable Supply in Uniform Price Auctions: Non-Commitment as a Strategic Tool". *Economics Letters*, April 2007.

McAfee, R. Preston and John McMillan, "Analyzing the Airwaves Auction," *The Journal of Economic Perspectives*, Vol. 10, No. 1 (Winter, 1996), pp. 159-175

Milgrom, Paul (2004), *Putting Auction Theory to Work,* Cambridge University Press

Milgrom, Paul (2008), "Assignment Auctions," working paper.

Mueller, Milton "New Zealand's Revolution in Spectrum Management," *Information Economics and Policy* 5, 2 (July 1993) 159-77.

PG&E (2007), Proposals for the Next Generation of Resource Adequacy of Pacific Gas and Electric Company: The Composite Approach & Multi-year Bilateral Resource Adequacy," California Public Utility Commission Rulemaking 05-12-013.

Salant, David, "Multi-Lot Auctions: Applications to Regulatory Restructuring." In *Obtaining the Best from Regulation and Competition*, edited by M.A. Crew and S. Spiegel. Boston, MA: Kluwer Academic Publishers (2004).