15 February 1999

Incentive-Compatible Evaluation and Settlement Rules: Multi-Dimensional Auctions for Procurement of Ancillary Services in Power Markets

Hung-po Chao, EPRI and Stanford
Robert Wilson, Stanford

Abstract: System operators in the electricity industry purchase reserve capacity in a procurement auction in which suppliers offer two-part bids, one part for making capacity available and another part for supplying incremental energy when called. Key ingredients of an efficient auction design are the scoring rule for comparing bids, and the settlement rule for paying accepted bids. We use the principle of incentive compatibility to establish that very simple rules suffice. In particular, the scoring rule uses only the capacity part of the bid, and energy supplies are paid the spot price. This is the design adopted in California.

1. Introduction

Procurement auctions in which bidders submit two-part bids are used in situations where each winning supplier provides two ingredients or two dimensions of service. Typically these auctions rely on a scoring rule to compare bids and a settlement rule to determine payments. Two-part bids are common in the electricity industry due to the distinction between an initial commitment of capacity availability and a subsequent determination of the amount of energy supplied from that capacity, contingent on later events.

Some notorious procurement auctions of this kind were California's 1993 series of 'Biennial Resource Planning Update' (BRPU) auctions. Each auction resembled a Vickrey auction in that the bidder with the lowest score won the right to negotiate a detailed supply contract on terms comparable to the bid with the second-lowest score. This feature was used to justify predictions that the auction outcome would be efficient, as in an ordinary auction under standard technical assumptions. In fact, however, the scoring and settlement rules encouraged bidding strategies that resulted in winning bids proposing huge upfront capacity payments and negative energy payments. (The BRPU auctions were eventually voided by the Federal Energy Commission on other grounds.)

Anomalous results of this kind stem from naive reliance on a scoring rule that uses a linear function of the two parts of a bid, and settles by paying the accepted bids.

The design of procurement auctions with two-part bids is a central problem in current efforts to restructure wholesale electricity markets. In particular, this problem arises in the auctions conducted by system operators (SOs) to purchase sufficient reserve capacity to meet contingencies. Thus, the two ingredients supplied by a winning bidder are capacity availability, and then depending on events, energy supply from that capacity when called by the SO during real-time operations. A typical example is (incremental) spinning reserve, in which a supplier operates a generator below its maximum rate so that it can be ramped up to higher rates when called to meet load surges. Besides the maximum operating rate of the generator, reliability standards restrict the amount of spinning reserve that a generator can provide to the amount that it can ramp to within a specified time, such as 10 minutes or 30 minutes. A typical ramp rate for a thermal generator is 1% of its maximum operating rate per minute, so at most 10% of capacity can be available within 10 minutes. Reliability standards are also the main determinant of the amount of spinning reserve purchased by the SO.

In California, the auction of spinning reserve is conducted roughly as follows. For spinning reserve to be provided in a specified hour of the next day, an auction is held after the close of the day-ahead forward markets for energy and transmission, which establish the amount of spinning reserve required by the SO to maintain reliability of the transmission grid. On Tuesday, for instance, 24 simultaneous auctions are conducted independently for reserves to be provided in the 24 hours of Wednesday. Each supplier who proposes to provide spinning reserve in a given hour submits a two-part bid for an amount that its ramp rate and maximum operating rate allow. Each bid specifies an offered price for capacity availability and an offered price for delivered energy. The SO then evaluates these bids using a scoring rule, and accepts enough bids to meet its reliability requirement. Those suppliers whose bids are accepted are obligated to maintain spinning reserve during the specified hour on Wednesday (enforced by spot checks and penalties for noncompliance). During this hour of real-time operations, the SO calls the spinning units as needed to meet load surges. The spinning units are called in order of increasing energy costs, called the *merit order*, as specified in the energy portion of their bids. The settlement rule specifies the payments made to the suppliers whose bids are accepted. These payments include a capacity payment derived from the initial evaluation, and an energy payment for the energy actually called and supplied.

This sketchy description of the market for spinning reserves omits various features that we ignore later. For instance, each bid is actually a schedule of quantities offered at different prices. Transmission constraints can alter the merit order. And there are other categories of reserves acquired in parallel auctions. Here, we use the auction of 10-minute spinning reserve as the context for our analysis.

We will, however, consider both incremental and decremental reserves. As in the description above, incremental reserves are provided by generators able to ramp up to higher production rates; similarly, decremental reserves are provided by generators able to ramp down to lower production rates in order to meet load decreases. Separate auctions are held for incremental and decremental reserves. The reasons for this important distinction will be explained in Section 4. (The energy part of a bid is usually unimportant for the reserve category called regulation, in which both increments and decrements are provided by 'automatic generation control' to follow the load continually, since such a unit is required to return to its set point every few minutes and therefore often has little or no net generation.)

Scoring and Settlement Rules

To indicate the role of scoring and settlement rules, we describe a version of the rules included in California's March 1997 filing to the Federal Energy Regulatory Commission, for the case of incremental spinning reserve. The bid format required each supplier to specify for each mega-Watt (MW) an offered capacity price, say R measured in \$/MW, and an offered energy price, say P measured in \$/MW-hour. If a supplier's bid were accepted then the settlement rule specified that the supplier would be paid $R + hP$ for that MW, where h is the fraction of the hour that its MW of energy was actually provided when called by the SO. The scoring rule was based on the SO's estimate, say H, of the expected fraction of the hour in which generation from spinning reserve would be needed during real-time operations. Thus, the SO assigned each bid (R, P) the score

$$I(R, P) \equiv R + H \times P .$$

Bids were to be accepted in increasing order of these scores until the SO's requirements for spinning reserve were satisfied. This scoring rule (the linear function I, depending on the parameter H) was justified by the objective of minimizing the SO's expected total cost of fulfilling its reserve requirements.

Problems with these scoring and settlement rules are evident from the previous experience with the 1993 BRPU auctions. Recall that during real-time operations the

spinning-reserve units are to be called in merit order; i.e., in increasing order of their energy bids. Each supplier can anticipate, therefore, that the duration h in which it is called will differ from the SO's overall estimate H — a scalar parameter that was not differentiated by unit or MW. Consequently, there are ample incentives for a supplier to 'game the system' by selecting the bid (R, P) to maximize its expected profit. For instance, based on an anticipated functional relationship $h(P)$ between its energy bid and the expected duration h that its unit is called, the supplier might choose its bid (R, P) to maximize the expectation of its realized profit $R - C + h(P)[P - c]$ based on its true fixed cost C and marginal cost c, taking account that the probability its bid is accepted depends only on its total score $R + HP$. For any specified score, this maximization could result in either a very high or very low energy bid P that is unrelated to its actual marginal cost c. This kind of gaming is similar to what occurred in the BRPU auctions.

Gaming of this sort distorts the energy portion P of a bid. Consequently, the SO's reliance on the merit order to call the spinning reserves undermines the efficiency of the outcome. That is, there is no assurance that the called energy is provided by those suppliers with the lowest marginal costs of generation.

This is a classic problem of incentives. What is needed are scoring and settlement rules that encourage each supplier to offer as the energy portion of its bid its actual marginal cost, and at the same time, enable the SO to procure the reserves it needs at the minimum expected total cost. Problems of this kind are addressed by the theory of mechanism design, in which procedural rules are constrained by 'incentive compatibility'. In the present context, this constraint requires truthful revelation, in the sense that if the procedural rules imply an optimal strategy of offering the energy price $P(c)$ when one's energy cost is c, then we require that $P(c) = c$. This constraint ensures that the merit order reflects suppliers' costs accurately.

In the following sections we examine the design problem in terms of standard results from mechanism design theory, and then derive scoring and settlement rules that implement an efficient design. Our method reverses the forward-looking approach of the 1997 design and instead works backward from the real-time market to the day-ahead procurement auction. Thus, our starting point is the scheduling of real-time calls in merit order: incentive compatibility constraints dictate the settlement rule for energy, which then enables a simple scoring rule for the procurement auction.

2. Design of the Real-Time Market

In California as in other jurisdictions, the economic role of reserves is to moderate the volatility of the real-time spot price. The SO ordinarily uses regulation and energy bids in the real-time market for following the load and for balancing the transmission grid, but when these bids are exhausted or unduly expensive it calls first on spinning reserves. When reserves are used, the energy bid of the last unit in the merit order among those called sets the spot price.

To keep matters simple, initially we consider only incremental reserves, so it is the most expensive among those units called that sets the spot price. In this case, the probability distribution of the spot price can be represented as follows. Let $F(Q)$ represent the probability that the quantity of generation called from incremental reserves will not exceed Q. Represent the merit order by the aggregate supply function $S(p)$, where S is a nondecreasing function of the price p. That is, $S(p)$ is the quantity of generation from those reserve units whose energy bids are less than the spot price p. Then the probability that the spot price will be less than p is

$$G(p) = F(S(p)) .$$

For instance, there is a high probability that the spot price will be less than p if there is a high probability that the supply available at price p exceeds the quantity called. Thus, $G(p)$ is the probability that the price required to meet load surges from the reserved capacity is less than p.

A basic assumption that we maintain throughout is that the distribution G of the spot price has a positive density. One part of this assumption is that the distribution F of quantities called has a positive density, so there is no lumpiness in the distribution of the SO's calls for energy. This does not assume that there is no chance of a large call, due perhaps to failure of a generator or a transmission line; rather, it assumes only that there is enough stochastic variability that each specific called quantity has negligible probability. This ensures that no supplier can design its energy bid to exploit some particular contingency. This depends on the time frame, of course, but since the procurement auction is held a day ahead this assumption is reasonable. Another part of this assumption is that the supply function S is smooth. This conveys the requirement that no supplier has a significant role in the aggregate supply. It is a strong assumption that incorporates the major feature of a competitive market, namely that each supplier's

offered capacity is small relative to the aggregate. This is more realistic when one takes account that each unit can offer only the quantity that it can ramp in 10 minutes.

This assumption already provides an indication of how the settlement rule can be constructed to meet the requirement for truthful revelation of each supplier's marginal cost via its energy bid. We present the construction first for a perfectly competitive market, and then provide a reinterpretation in terms of a Vickrey auction.

Revelation in a Perfectly Competitive Spot Market

In a competitive market, each supplier has a negligible chance of affecting the spot price. Therefore, if the settlement rule merely pays the spot price for called energy, then a supplier's optimal bid — conditional on winning in the day-ahead auction — is simply to name its marginal cost as its energy bid. To see this, re-interpret the energy bid P as the supplier's reserve price, namely the least spot price at which it wants to be called. Then the expected profit of a supplier whose marginal cost is c is

$$\Pi(P, c) = [1 - G(P)] \times E[p - c \mid p \geq P] .$$

This formulation reflects the fact that the supplier's unit is called whenever the spot price p exceeds its reserve price P, so the probability of being called is $1 - G(P)$, and when it is called its expected profit is the conditional expectation of the profit margin $p - c$. When G has a positive density at $p = c$, as assumed, the unique maximizer $P(c)$ of this expected profit is $P(c) = c$, as required.

This settlement rule implies that the expected profit from called energy for a supplier with marginal cost c is $\Pi(c, c)$, which is a convex and decreasing function of c.

Revelation in a Vickrey Auction

An analogous result is obtained from a Vickrey auction. In the simplest case where each supplier provides a single MW, the settlement rule specifies that called energy is paid the lowest rejected energy bid in the merit order. Now the probability distribution G has the analogous interpretation that $G(p)$ is the probability that the lowest rejected bid is no more than p. As before, the expected profit is

$$\Pi(P, c) = [1 - G(P)] \times E[p - c \mid p > P]$$

and again the optimal bid $P(c) = c$ reveals the supplier's marginal cost accurately.

More generally, a Vickrey auction requires that the q-th MW provided by a supplier is paid the q-th from the lowest energy bid (one for each MW) among those rejected that are not bids from this supplier. This settlement rule is more complicated, and in particular some large suppliers may be paid more than the nominal spot price, but it has the advantage once again that, if suppliers' costs are statistically independent, then an optimal strategy is to specify one's marginal cost for each MW as the reserve price, namely $P(c) = c$.

The settlement rule for a Vickrey auction ensures truthful revelation of marginal costs even if the energy market is imperfectly competitive. In the sequel, however, we concentrate on the perfectly competitive case in order to clarify the derivation of the scoring rule used in the initial procurement auction. Thus, we use $\Pi(c, c)$ as the formula for a supplier's expected profit in the energy market even though this is not exactly accurate for a Vickrey auction when competition is imperfect.

Interpretation

For both the competitive spot market and the Vickrey auction, the above derivations are incomplete because in each case the analysis is conditioned on the supplier having won a place in today's merit order when it bid in the reserve auction conducted yesterday. The result says only that, for a settlement rule that pays the spot price for energy, *after* winning spinning reserve status in the day-ahead auction, a supplier's optimal reserve price for energy is its true marginal cost, namely $P(c) = c$. This leaves open the question of whether the supplier might prefer to distort its day-ahead energy bid to improve its chances of winning reserve status. The resolution of this question depends on the scoring rule, which we examine next.

3. Design of the Procurement Auction

A brief summary of Section 2 is a sequence of implications. Productive efficiency of the real-time energy market requires calling the reserved units in the merit order of their marginal costs. A settlement rule that pays the spot price for called energy suffices to encourage each supplier to offer its actual marginal cost as its energy bid, interpreted as a reserve price below which it prefers not to be called — provided this does not alter the chances of being selected to provide spinning reserve. Therefore, a procurement auction that ignores suppliers' energy bids in selecting those awarded reserve status solves the problem of productive efficiency in the next day's energy market.

One such auction is a Vickrey auction in which those suppliers offering the lowest capacity prices are accepted. That is, the scoring rule is simply

$$I(R, P) = R .$$

Again, the settlement rule of a Vickrey auction pays for the q-th MW of spinning capacity accepted from a supplier the q-th lowest rejected bid (one for each MW) among those from *other* suppliers.

To simplify, however, we again invoke the assumption of perfect competition. In this case every accepted MW is paid the same price, say R^*, which is the capacity price offered by the lowest rejected bid. In sum, the settlement rule in the perfectly competitive case is that each MW accepted in the procurement auction is paid R^* for making spinning capacity available, plus the spot price p for energy that is actually called and supplied.

Our aim in the remainder of this section and the next is to establish that such an auction in the day-ahead market for reserved capacity is efficient for a standard specification of the economics of supply for reserve capacity. In this specification, the costs incurred by a supplier providing spinning reserve are independent of its energy bid P. For example, a supplier might incur a direct cost maintaining a unit in spinning condition, and further, an opportunity cost representing the expected profit that could otherwise have been earned if the reserved capacity were instead committed to producing energy for sale in the day-ahead and/or real-time energy markets. The latter may be problematic if the supplier's real-time sales are large, but here we assume that, as in competitive spot markets, a supplier's energy bid P has a negligible chance of affecting the spot price p in the real-time market. Thus, we assume that a supplier with marginal cost c has a reserve price in the procurement auction that is the difference

$$R(P, c) = V(c) - \Pi(P, c)$$

between its foregone expected profit $V(c)$ and its expected profit $\Pi(P, c)$ from called energy paid the spot price.

For instance, if the opportunity cost is the profit that could have been earned in the day-ahead energy market where the clearing price is π, then $V(c) = \max\{0, \pi - c\}$ independently of the energy bid P offered in the auction of spinning reserve. As in this example, we assume that V is a nonincreasing function of the marginal cost c.

The Scoring Rule

To examine the effect of the scoring rule, it suffices to illustrate the case that it has the additive form

$$I(R, P) = R + H(P),$$

where (as in the 1997 filing) H is a smooth nondecreasing function of the energy part of the bid; and, the settlement rule pays each accepted unit the capacity price

$$R^*(P) = I^* - H(P),$$

where I^* is the smallest score among those rejected. Thus, if the bid (R, P) is accepted then that supplier is paid $R^*(P)$ for reserving capacity, plus the subsequent spot price p for energy called when $p > P$.

We claim that in this case the incentive compatibility constraint requires that $H(P)$ does not depend on the energy bid P; that is, it is *necessary* for productive efficiency and the validity of the merit order that bids are evaluated solely on the basis of the capacity part of each bid. To show this requires several steps.

1. Observe first that a supplier's bid (R, P) will be accepted if and only if $I(R, P) < I^*$, where I^* does not depend on its bid, and if accepted its expected profit will be $R^*(P) - R(P, c)$ net of any opportunity and fixed operating costs. For any fixed choice of P, therefore, the bidder's incentive is to offer the capacity price $R = R(P, c)$, since that maximizes the probability its bid is accepted without altering its expected profit if accepted.
2. The next step observes that the expected profit from an accepted bid is

   $$R^*(P) - R(P, c) = I^* - H(P) - V(c) + \Pi(P, c).$$

   The incentive compatibility constraint that the optimal energy bid must be $P = c$ therefore requires that

   $$H'(c) - \Pi_1(c, c) = 0.$$

   Note further that this constraint reinforces the optimality of the choice $R = R(P, c)$ for the capacity bid, since $R(P, c) = V(c) - \Pi(P, c)$ is minimized by choosing $P = c$ as we saw in Section 2.
3. The third step uses the fact that $\Pi_1(c, c) = 0$ to infer that $H'(c) = 0$, which implies that H must be a constant.

This derivation can be extended to conclude that the SO's optimal choice of the constant H is actually $H = 0$. That is, a negative value increases I^*, providing unnecessary payments to the winning bidders, and a positive value decreases I^*, thereby unnecessarily excluding some bidders.

A similar construction applies to decremental reserves. In this case a winning bidder's expected profit from real-time calls from the SO is

$$\Pi(P, c) = G(P) \times E[c - p \mid p < P] .$$

This formula reflects the fact that for a decrement the supplier pays the spot price to the SO to purchase energy that replaces its previous delivery commitments contracted in the day-ahead energy market, for which it would otherwise incur the marginal cost c. The opportunity cost is typically nil, $V(c) = 0$, in view of the supplier's previous opportunity to sell less in the day-ahead energy market.

Example: Suppose that the spot price p in the real-time market is distributed uniformly on the interval $[\pi - \Delta, \pi + \Delta]$. Then for incremental spinning reserve,

$$\Pi(P, c) = \frac{1}{4\Delta}[(\pi + \Delta - c)^2 - (P - c)^2] ,$$

provided $P \leq \pi + \Delta$. The optimal bid offers the energy price $P = c$ and the capacity price

$$R = R(c, c) = V(c) - \Pi(c, c) .$$

In particular, if π represents the clearing price in the day-ahead energy market and $V(c) = \max\{0, \pi - c\}$, then

$$R = \begin{cases} -(1/4\Delta)(\pi - \Delta - c)^2 & \text{if } c \leq \pi , \\ -(1/4\Delta)(\pi + \Delta - c)^2 & \text{if } c \geq \pi . \end{cases}$$

Notice that the capacity bid R is negative in this example, reflecting the fact that reserve status is essentially a call option exercised whenever the spot price exceeds marginal cost. Indeed, in actual operation, procurement auctions for spinning reserve sometimes clear at zero prices, or if allowed (as in California), at negative prices. The clearing price is usually positive, however, because maintenance of spinning reserve incurs fixed costs, and for units that provide only spinning reserve there are also start-up costs, possibly augmented by unprofitable sales of energy from generation at the minimum feasible rate.

4. Overall Productive Efficiency

Lastly we derive conditions under which this auction design provides an efficient schedule of overall capacity and energy purchases. The traditional analysis relies on two ingredients. One is the load-duration curve, whose role is played here by the probability

distribution F of the amount of called energy. The other is the set of efficient technologies, whose role is played here by the locus of pairs in which each pair $(V(c), c)$ represents the fixed cost $V(c)$ of reserving capacity whose marginal cost is c.

The Locus of Efficient Technologies

In the general case, a production technology with fixed cost C and marginal cost c is efficient if there is a duration h for which the total cost $C + hc$ is minimized by this technology compared to any alternative. In our case, the assumption that $V(c)$ is a nonincreasing function is insufficient in itself to assure that every pair $(V(c), c)$ is on the efficient locus. Where $V(c)$ is strictly decreasing the pair $(V(c), c)$ is on the efficient locus. But if $V(c)$ is the same for each marginal cost in an interval $c_* < c < c^*$ then only the pair $(V(c_*), c_*)$ is on the locus of efficient technologies. Thus, it is necessary for productive efficiency to exclude those technologies with the same fixed cost and higher marginal costs.

This consideration implies that a particular tie-breaking rule should be used in the procurement auction: if two bids offer the same capacity price then the one with the lower reserve price for called energy has priority for acceptance. In practice this tie-breaking rule may have few effects because each bid for 10-minute spinning reserve offers a small quantity, so even if one with lower marginal cost is accepted first it is still necessary to accept the second too in order to meet the SO's demand for reserves. However, for the particular formula $V(c) = \max\{0, \pi - c\}$ that is commonly used in examples to illustrate markets for incremental reserve, the tie-breaking rule might play a more significant role. All suppliers with costs in the upper range $c > \pi$ of marginal costs exceeding the day-ahead price π have no opportunities to sell in the day-ahead energy market and therefore all have the same opportunity cost $V(c) = 0$, and therefore each offers the capacity bid $-\Pi(c, c)$ depending on its marginal cost c. In this case, a lower marginal cost c is associated with a higher profit $\Pi(c, c)$ from called energy, and so indeed the capacity bid $R(c, c) = -\Pi(c, c)$ is lower and therefore is accepted first. More generally, however, such suppliers might differ in other cost ingredients, such as start-up or no-load costs required to make capacity available, so the tie-breaking rule can play a role in selecting between two or more with the same capacity bids and differing energy bids by giving priority to the one with the smaller energy bid.

The tie-breaking rule is especially important in those systems that constrain bids to be nonnegative, because then there are often several capacity bids that are zero. These

same considerations apply to decremental reserve, since again the opportunity cost is often supposed to be nil, so the tie-breaking rule can be a major factor in selecting among identical capacity bids by assigning priority to those offering the highest energy bids for called decrements.

It is important to realize, however, that constraining capacity bids to be nonnegative should be avoided to prevent 'gaming the system'. The chief risk is that an inefficient generator (one for which $C + hc$ is not minimal for any duration h within the hour) could bid a positive capacity price that is always accepted by the SO and an energy price so high that it is never called for generation: in this case the supplier collects capacity payments without providing any useful service. This gaming strategy is precluded if the capacity payment is negative (i.e., the supplier pays the SO for reserve status) or less than the start-up and no-load cost of making the unit available — provided of course that, as in California, the SO occasionally checks that reserved capacity is in fact operable and responsive to calls for incremental generation.

The Efficient Production Schedule

The load-duration curve is usually constructed by recording for each load Q the number or fraction $H(Q)$ of hours in the year that the load is at least Q; or in reverse, by recording the load $Q(H)$ of the H-th hour when the hours are ranked in order of their loads. The corresponding representation in the present context is the probability $1-F(Q)$ that the real-time call for energy exceeds Q, where this probability is interpreted as the expected fraction of the hour of real-time operations.

The traditional analysis selects the lowest-cost technology, say (C^*, c^*), to serve the peak load of short duration H^*; thus C^* is typically small and c^* is typically large, which represents an efficient tradeoff because the capacity is idle most of the time. Then all capacity is paid the 'demand charge' C^* plus the spot price for delivered energy. A technology with a higher capacity cost but lower marginal cost, say (C, c) that is most efficient for the longer duration $H > H^*$ is called whenever the load exceeds $Q(H)$. This corresponds to the familiar operating rule that calls the efficient technologies in the merit order of their marginal costs as the load rises over an interval of real-time operations. The technology (C, c) recovers its total cost from the demand charge C^* and the spot price that is the marginal cost of the last unit called in merit order. In particular, even though its capacity cost C exceeds the demand charge C^*, its lower marginal cost $c < c^*$ enables a baseload unit to recover its full cost from longer profitable

runs at energy prices exceeding its marginal cost. [In the ideal world of the mathematical model this cost recovery is exact when each technology has constant returns to scale: compared to a peaking unit, the baseload unit's greater profits from energy sales exactly compensate for its higher capacity cost, so $C - \Pi(c, c) = C^* - \Pi(c^*, c^*)$. With limited capacity available from each technology, however, there may be additional scarcity rents that can be earned so these differences can be negative, as in the example above.]

This traditional analysis applies equally to the procurement auction. The load-duration curve is replaced by the probability $1 - F(Q)$ that a quantity exceeding Q is called in the hour of real-time operations; and the set of efficient technologies is replaced by the pairs $(V(c), c)$ of suppliers' fixed costs of reserving capacity and marginal costs of generation, where $V(c)$ is a decreasing function of c. The merit order remains the same, since paying the spot price for energy encourages each bidder to offer $P(c) = c$ as its energy bid.

There is one very significant difference, however. To describe and analyze this difference we rely on the standard example in which there are no fixed costs and the opportunity costs are $V(c) = \max\{0, \pi - c\}$ for incremental reserve and $V(c) = 0$ for decremental reserve, where π is the clearing price in the day-ahead energy market. To distinguish between these two reserve categories, let $R_I(c) = V(c) - \Pi(c, c)$ be the predicted capacity bid in the auction for incremental capacity, and similarly $R_D(c)$ is the predicted capacity bid for decremental capacity.

The relevant fact is that for incremental reserve the predicted capacity bid $R_I(c)$ is not an increasing function of the marginal cost c as required for full efficiency. Indeed, $R_I(c)$ decreases with slope $-G(c)$ for $c < \pi$ and increases with slope $1 - G(c)$ for $c > \pi$. Consequently, an auction in which suppliers with costs on both sides of π participate will typically accept some from each side. This non-monotonicity undermines the efficiency of the outcome that would otherwise be predicted by the theory of mechanism design. But apart from this mathematical problem it is already evident that acceptance of bids from suppliers with costs below π violates the intent of the example's formulation. That is, attracting low-cost suppliers away from the day-ahead energy market is inefficient because energy sales in that market are sure whereas energy sales from reserved capacity are uncertain. Efficiency therefore requires that the auction design not attract away from the energy market those suppliers who can successfully sell there at the clearing price π.

To resolve this problem we examine the role of decremental reserves. A simple

calculation shows that $R_D(c)$ decreases with slope $-G(c)$ in the range $c < \pi$, just as $R_I(c)$ does. Therefore, either these two functions are the same in this range or one dominates the other. For efficiency, we require that $R_I(c) \geq R_D(c)$ so that those suppliers with low costs $c < \pi$ prefer to sell in the day-ahead energy market at the price π and then offer bids for decremental reserve (rather than incremental reserve). There is, in fact, a simple and realistic assumption that ensures that this requirement is satisfied:

$$E[p] = \pi .$$

That is, if the day-ahead price π is an unbiased predictor of the real-time spot price p then $R_I(c) = R_D(c)$ for every marginal cost $c \leq \pi$. This assumption is just the no-arbitrage condition that one can realistically suppose to characterize a sequence of markets. For example, if $E[p] < \pi$ then suppliers would prefer to defer sales from the day-ahead market to the real-time market.

A touch of realism strengthens this result. Decremental reserve requires no fixed costs, whereas incremental reserve incurs no-load costs and possibly start-up costs, so typically the no-arbitrage condition $E[p] = \pi$ actually implies the strong inequality $R_I(c) > R_D(c)$ when $c < \pi$. We interpret this resolution of the problem as a basic explanation for the separate auctions of incremental and decremental reserves conducted by system operators. It conforms, moreover, to the conventional wisdom that all supply units that can sell in the day-ahead market do so and offer bids only for decremental reserve, whereas it is only those whose costs are too high to sell in the day-ahead market who offer bids for incremental reserve.

5. Concluding Remarks

We have shown that very elementary considerations can be used to derive an efficient design for a procurement auction in which the system operator purchases reserve capacity that it can later call for energy generation. The key to our approach is to work backward from the requirement that, for productive efficiency, the spinning reserve units must be called in merit order based on marginal costs that are accurately revealed by supplier's bids in the initial procurement auction.

The distinguishing features of such auctions are two-part bids, one part offering a price for capacity availability and another offering a reserve price for energy called in real-time. Because efficient generation requires the merit order to reflect accurately the suppliers' marginal costs, we impose the incentive compatibility condition that each

supplier's optimal energy bid must accurately reveal its privately known marginal cost. One way to ensure this condition is to use a variant of the Vickrey auction in the spot market. In the perfectly competitive case, each energy bid is interpreted as a reserve price to construct the merit order, and then all energy is paid the spot price obtained as the lowest unused energy bid. To ensure that incentive compatibility is not distorted by the scoring rule used to compare bids in the initial reserve auction, it is necessary that only the capacity prices are used in the scoring rule for comparing bids. This ensures that the system operator obtains the required reserves at least cost, since those bids accepted are the ones for which the difference between the fixed and opportunity costs of reserving capacity and the expected profit from spot market sales is smallest. A Vickrey auction can be used at the initial stage, or in the competitive case, all winning bidders are paid the lowest rejected bid offered for reserving capacity. This payment corresponds exactly to the demand charge familiar from traditional analyses of efficient capacity planning, since it represents the fixed and opportunity costs of the last (peaking) unit in the merit order among those accepted for reserve status, net of profits from called energy paid the spot price.

Overall productive efficiency depends crucially on two additional features. One is that parallel auctions for incremental and decremental reserves are conducted, and the second is that arbitrage between the day-ahead and real-time markets is sufficient to ensure that the day-ahead energy price is an unbiased predictor of the spot price, namely $E[p] = \pi$. These features ensure that low-cost suppliers prefer to sell energy in the day-ahead market, rather than providing incremental reserves, and offer bids only in the auction for decremental reserves.