

A Review of ISO New England's Proposed Market Rules

Peter Cramton and Robert Wilson¹
Market Design Inc.

September 9, 1998

Executive Summary

This report reviews the proposed rules for restructured wholesale electricity markets in New England. We review the market rules, both individually and collectively, and identify potential problems that might limit the efficiency of these markets. We examine alternatives and identify the key tradeoffs among alternative designs. We focus primarily on questions raised in our Scope of Work.

We believe that the wholesale electricity market in New England can begin on December 1, 1998. However, improvements are needed for long-run success. We have identified four major recommendations:

- Switch to a multi-settlement system.
- Introduce demand-side bidding.
- Adopt location-based transmission congestion pricing, especially for the import/export interfaces.
- Fix the pricing of the ten minute spinning reserves.

Of these, only the final (and least important) recommendation realistically could be implemented by the December 1st start date. We do not view this as a fatal problem, provided that by the start date the ISO and NEPOOL reach agreement in principle on these basic concepts and a tentative timetable for implementation. We believe that if the markets open without any sense of what improvements will be made or when they will be made, then it will be much more difficult to adopt and implement needed improvements. An evolutionary “wait and see” approach would be too slow, and likely would result in damage to the markets that is difficult to correct.

Our analysis is broad and abstracts from many of the details of implementation. Because of the level of abstraction, our recommendations must be viewed as tentative. It will be important for ISO New England to monitor closely the markets in the early months to identify and correct problems as they appear.

¹ Cramton: cramton@umd.edu, (301) 405-6987; Wilson: rwilson@stanford.edu, (650) 723-8620. This report was commissioned by ISO New England. The views expressed are our own. © Market Design Inc. 1998.

Our recommendations stem from the following critique of the proposed market rules:

◆ **The single-settlement system is vulnerable.**

The single-settlement system creates strong incentives to manipulate the spot price and reschedule. This is easy for participants to do through short-notice imports/exports and scheduling changes. There is sufficient market power in New England to make it likely that spot price manipulation will be significant. The absence of congestion pricing allows exploitation of imports/exports and other means as the device. This is complicated by inconsistencies with neighboring jurisdictions' methods of managing transmission congestion, and further complicated by arcane rules for determining the spot price (reliance on forward-rolling LPs or other methods, none of which recognize the full value of flexible resources). We expect the ISO to encounter severe problems trying to establish reliable, stable, feasible scheduling—the parties will manipulate schedules essentially every day and every hour, and try constantly to use short-notice imports/exports with NY/Quebec to the extent transmission allows.

◆ **The spot-price system is problematic.**

The 5-min LP optimization, in terms of treatment of intertemporal constraints and allowance for forecast errors and other contingencies, is fundamentally inconsistent with the day-ahead optimization. Participants are not being paid real-time for the commitments made day-ahead, and peak prices are biased too low, which disadvantages flexible resources. The absence of demand-side bidding both day-ahead and spot makes the spot price biased upward.

◆ **The absence of location-based congestion pricing is inefficient.**

The absence of transmission congestion pricing, especially for imports/exports, is fundamentally inefficient, and likely to engender ferocious arbitrage and manipulation vis-a-vis NY/Quebec/New Brunswick. Paying out-of-merit-order generators their bids invites gaming. It is better to pay them and every one else in a congestion zone the marginal bid that is scheduled. The proposed pricing will induce manipulation of day-ahead supply functions in a game to garner constrained on/off payments. Moreover, the current rules fail to provide the correct incentives for the location of new generation.

◆ **The viability of the day-ahead energy market is not obvious.**

The existing structure encourages self-scheduled bilateral contracts to take over the New England energy market for forward transactions, leaving the ISO to manage real-time balancing and reserves. We are concerned that if the ISO energy market has a tiny share of the total market then physical feasibility will be more precarious, real-time balancing will be more difficult, and the spot price will be more volatile. This will have a negative impact on the efficiency of bilateral contracting and the

electricity market as a whole. The physical feasibility of self-scheduled bilateral transactions is problematic, in part because the incentives encourage reliance on the ISO to remedy difficulties in the real-time market.

We list our recommendations (four main recommendations and four secondary recommendations) roughly in order of importance.

1. ISO New England should switch to a multi-settlement system. Although it could not be implemented by December 1, its implementation is helped by the fact that the neighboring power markets in PJM and NY have adopted a multi-settlement system. We recommend either a two-settlement system with financially binding day-ahead bids, or a three-settlement system with binding bids both day-ahead and hour-ahead. Such an approach provides market incentives for participants to respond efficiently to uncertain demand and supply. Deviations from day-ahead schedules are priced by the market. Moreover, it mitigates incentives for gaming and reduces bidder uncertainty. Adopting a multi-settlement system would eliminate the inherent gaming problems that short-notice transactions create. Multiple settlements, self-scheduling, and day-ahead commitments are natural organizational complements that cannot be implemented piecemeal.
2. Demand-side bidding should be introduced as soon as possible. The ISO should develop the rules and other steps needed for implementation, taking advantage of the experience of other markets such as California. Demand-side bidding will mitigate supplier market power and provide incentives for power management. Demand-side bidding is essential for long-run efficiency.
3. The uniform uplift approach to transmission congestion is unlikely to be adequate. The ISO should begin investigating alternatives. Location-based pricing will increase short-run efficiency, especially regarding imports/exports used to arbitrage prices between New England and neighboring jurisdictions, and provide improved incentives for generation and transmission siting and expansion. Also it will improve the stability of the spot price.
4. The pricing of spinning reserves should be greatly simplified. At the very least, the double-counting should be eliminated. A preferred approach would be to let all capable resources bid for TMSR and other reserves in a cascade.
5. The risk of implicit collusion is sufficiently large to outweigh the efficiency gains of disclosing information beyond market prices and total quantities. The ISO should have in place before the market opens a plan to analyze the bidding data on a continuing basis for signs of the use of market power.
6. The ISO and NEPOOL should consider eliminating the installed capability market and the operable capability market. In the long run, incentives for sufficient capacity are provided by the energy and

reserve markets. If the installed capability market is retained, it may be desirable to allow iterative bidding. The installed capability market does not appear to be an effective deterrent to over-expansion of capacity, and the operable capability market appears ineffective in avoiding gaming of maintenance schedules.

7. Only pure energy bids are needed for efficiency in the energy market. The inclusion of start-up and no-load components, while sensible in a centrally optimized system, are prone to gaming in a decentralized bidding environment. Bidders should be able to self-schedule their units reliably through pure energy bids, especially under a multi-settlement system.
8. Iterative bidding in the energy market is probably unnecessary, especially if the ISO switches to a three-settlement system. Then the bidders will have an opportunity to correct scheduling deficiencies in the hour-ahead and spot markets.

1 Introduction

This report reviews the proposed rules for restructured wholesale electricity markets in New England. The markets are auction markets. We review the auction designs, both individually and collectively, and identify potential problems that might limit the efficiency of these markets. We examine alternatives and identify the key tradeoffs among alternative designs. We conclude the report with several recommendations. Most of these are long-term recommendations, in the sense that they could not be implemented before the December 1st start date.

Our analysis is broad and abstracts from many of the details of implementation. This approach is required due to the tight time constraints. Also, we believe it is the best approach as ISO New England plans for the initiation and future development of a competitive electricity market. However, because of the level of abstraction, our recommendations must be viewed as tentative. Some of the recommendations undoubtedly will change as we learn more about competitive electricity markets, and the New England market in particular. Moreover, there certainly will be implementation problems that we are unable to anticipate. Thus, it will be important for ISO New England to monitor closely the markets in the early months to identify and correct problems as they appear.²

The issues that we address here concern the organization of the wholesale markets for energy and transmission, interpreted as including ancillary services and other requirements for system reliability and security. The examination of these issues in New England can benefit from the history of restructuring in other countries, such as Britain, Australia, New Zealand, Norway, Spain, Canada, and current

² The implementation problems encountered in the early months of operation of the California and PJM systems are worth reviewing as indications of what ISO New England might expect.

developments in the U.S., especially California, New York, and PJM. We emphasize the implications of the general principles of market design based on ideas from economics and game theory.

The special features of the electricity industry that must be considered include temporal and stochastic variability of demands and supplies, accentuated by the non-storability of power, multiple technologies with varying sensitivities to capital and fuel costs and environmental and siting restrictions, and dependence on a reliable and secure transmission system. The economic problems include substantial non-convexities (immobility of generation and transmission facilities, scale economies in generation, non-linearities in transmission), and externalities (mainly in transmission). As regards generation these problems have eased sufficiently in recent decades to enable competitive energy markets, but they remain important considerations in designing these markets.

The criteria for selecting among market designs include efficiency over the long term, including incentives for investment in facilities for generation and transmission. However, our exposition focuses on short-term efficiency, since this is the immediate concrete problem, and it is required for long-term efficiency.

Our perspective emphasizes strategic behavior. This seems paradoxical, since our aim is to construct a design that suppresses gaming or renders it ineffective in favor of greater efficiency. The principle, however, is to treat the market design as establishing a mode of competition among the traders. The key is to select a mode of competition that is most effective in realizing the potential gains from trade.

We begin in Section 2 with a brief description of the ISO-managed markets in New England, the likely market structure when operations begin, and the properties of efficient market rules. The subsequent sections address the topics specified in our Scope of Work. Sections 3-6 discuss four major issues that the rules must address: the settlement system, transmission congestion, the interaction between the energy and the reserve markets, and revelation of bidding information. Sections 7-9 address issues in the specific markets. Section 10 concludes with a recap of our perspective.

2 Background

2.1 Description of the ISO's Markets

Seven markets are proposed: (1) the energy market, (2-5) four markets for ancillary services, and (6-7) two capacity markets. A brief description of each is given below.

1. The *energy market* is a residual market. Only the difference between a participant's energy resources and its energy obligations is traded in the ISO market. These resources and obligations include amounts covered by bilateral contracts. Hourly bids, expressed in \$/MWh, are submitted

on a day-ahead basis for the next 24 hours. The ISO then schedules the generating units that will run the following day based on minimizing total costs in the energy market, as represented by the accepted bids. The market is settled after the fact on an hourly basis. All transactions are priced at the (ex post) energy clearing price. Payments/receipts are equal to the MWh bought/sold times the clearing price. Suppliers are paid for out-of-merit-order dispatch to alleviate transmission congestion on the basis of their bids submitted in the energy market.³

2. The *ten minute spinning reserve (TMSR) market* is a full requirements market. All TMSR is bought/sold through the ISO. Bidding and settlement are done as in the energy market—hourly bids in \$/MW for the next day are submitted, and the markets are settled hourly after the fact. Given the units dispatched to provide energy, the ISO selects the least-cost resources to provide required TMSR, taking into account bid costs, lost opportunity costs, and production cost changes. When the market begins operation, only hydro units and dispatchable loads may bid into this market. Although they cannot bid, fossil-fueled generators can be selected to participate in the market based on lost opportunity costs and production cost changes. Designated resources are paid the energy clearing price for any MWh provided *plus* lost opportunity cost *plus* production cost changes *plus* the bid times the MW provided. The total cost of providing TMSR is shared proportionally by load.
3. The *ten minute non-spinning reserve (TMNSR) market* is a full requirements market. All TMNSR is bought/sold through the ISO. Bidding and settlement are done as in the energy market—hourly bids in \$/MW for the next day are submitted, and the markets are settled hourly after the fact. Designated resources are paid the clearing price times the MW provided as reserved capacity. The total cost of providing TMNSR is shared proportionally by load.
4. The *thirty minute operating reserve (TMOR) market* is a full requirements market. All TMOR is bought/sold through the ISO. Bidding and settlement are done as in the energy market—hourly bids in \$/MW for the next day are submitted, and the markets are settled hourly after the fact. Designated resources are paid the clearing price times the MW provided. The total cost of providing TMOR is shared proportionally by load.
5. The *automatic generation control (AGC) market* is a full requirements market. All AGC is bought/sold through the ISO. Bidding and settlement are done as in the energy market—hourly bids for the next day are submitted, and the markets are settled hourly after the fact. AGC is measured in *regs*, which measures a unit's ability to follow load. Units that can provide AGC at lowest cost based on bids, lost opportunity costs, and production cost changes are selected.

³ The bidder may not be paid its bid if certain transmission congestion structure and price screens are not satisfied.

Generators providing AGC are paid the clearing price for time on AGC times the number of regs *plus* a payment for AGC service actually provided *plus* any lost opportunity cost. The total cost of providing AGC is shared proportionally by load.

6. The *operable capability market* is a residual market. Only the difference between a participant's operable capability resources and its operating capability obligation (load plus operating reserve) is traded through the ISO. Bidding and settlement are done as in the energy market—hourly bids in \$/MW for the next day are submitted, and the markets are settled hourly after the fact. A clearing price is calculated based on the bids of those participants with excess operable capacity. Participants who are deficient in operable capability pay the clearing price for each MW to those who are in surplus and who bid a price less than or equal to the clearing price.
7. The *installed capability market* is a residual market. Only the difference between a participant's installed capability resources and its installed capability obligation (load plus installed operating reserve) is traded through the ISO. Trading in this market occurs monthly. Bids are submitted in \$/MW-month on the last day before the month begins. A clearing price is calculated based on the bids of those participants with excess installed capacity. Participants who are deficient in installed capability pay the clearing price for each MW-month to those who are in surplus and who bid a price less than or equal to the clearing price. FERC currently has a cap of \$8,750/MW-month.

The energy and reserve markets (1-5) represent the core of the system; there is no separate market for transmission. These markets cannot be viewed in isolation, as each interacts with the others. In what follows, we examine the markets as a system, recognizing any interdependencies.

2.2 Market Structure in New England

Full efficiency requires competitive markets—markets in which no single participant can significantly influence prices. Due to economies of scale and transmission constraints, electricity markets are unlikely to achieve the ideal of perfect competition. Some parties will have some market power in at least some local markets (load pockets), a feature reinforced by payment of bids for out-of-merit dispatch to alleviate congestion.⁴ As a result, the electricity market structure in New England will be relevant in assessing the market rules, and determining how robust the rules are to the reality of market power. Recommendations will depend in part on the extent of market power.

⁴ As mentioned in footnote 3, there are price and market structure constraints limiting when a bidder may be paid its bid for out-of-merit dispatch.

There are three interdependent determinants of market power: market shares, transmission congestion, and market contestability. For the most part, we will assume here that transmission congestion is not a problem in New England. Transmission congestion will be addressed separately in Section 4.

Our discussion of market structure is based on the market shares of generation in New England as it is likely to be at the time the markets begin operation. Assuming transmission congestion is not a problem, market power is likely to be an issue only with respect to the two largest participants, NU and PGET. These two bidders have over 50% of the market's bidding authority and operating authority, ignoring imports. Because of joint ownership of some units, NU and PGET have knowledge of, and influence on, an even larger share of the market. In our mind, these market shares may be a cause for concern. Also, market power may be a more important issue if transmission congestion is a problem in some locations at peak times.

In evaluating market power in the various auction markets, one should look at the market shares conditional on a unit's ability to participate in the relevant market. For example, only hydro units and dispatchable loads can bid into the TMSR market; only units capable of providing AGC can participate in the AGC market. In addition, baseload units at the bottom of the supply curve, such as nuclear units or other units with low operating costs that are unlikely to set the clearing price at peak times, can be excluded from the market share calculations. What is relevant in evaluating market power is the shape of the supply curve around the market clearing price; that is, the elasticity of supply at the margin, especially in peak periods. If the supply curve is steep, then the larger bidders, will be able to influence the clearing price, and will have a strong incentive to reduce the quantity offered to achieve a higher clearing price.⁵

Specific recommendations on market power mitigation are presented in Section 6.

2.3 Properties of Efficient Market Rules

In evaluating the market design, we take the objective to be *to promote an efficient market for wholesale electricity in New England*. We now discuss several key properties of efficient market rules.

◆ Do the rules send the right price signals?

The pricing rules are critical in assessing efficiency in markets. Prices provide the incentives that influence behavior. For full efficiency, prices should reflect the marginal costs of suppliers and the marginal values of demanders.

⁵ The economics of quantity reduction and the inefficiencies it creates are studied in Ausubel and Cramton (1998). This behavior is observed in several experimental studies, as well as in the real markets. For experimental evidence,

◆ **Do the rules minimize opportunities for gaming?**

Gaming arises from several sources. Incorrect price signals give bidders a direct incentive to act in a way contrary to efficiency. Market power enables large bidders to distort prices in their favor. Non-binding bids, unpenalized scheduling changes, and weak monitoring and enforcement of compliance, encourage misrepresentation.

◆ **Do the rules enable markets to be contested?**

Contestable markets discipline incumbents via the threat of entry. To the extent that entry is swift and costless, incumbents are unable to exercise market power. Efficient rules encourage entry. Rules with low bidding costs and few bidding restrictions encourage entry. Simple and transparent rules are often best since they tend to minimize the cost of participation. Such rules encourage participation by small bidders and entrants that can undermine attempts by large incumbents to exercise market power.

◆ **Are the rules neutral with respect to bilateral transactions?**

Efficient rules let the bidders decide how best to participate in the market. Bilateral transactions are neither encouraged nor discouraged. Innovative and customized long-term contracts enable realization of mutual advantages that are not priced explicitly in the general markets.

◆ **Do the rules mitigate opportunities for collusive behavior?**

The essential ingredients to mitigate collusive behavior are: (1) to limit market power, (2) to encourage entry, and (3) to limit participants' ability to coordinate and enforce collusive outcomes.

We use these criteria as guides in our examination of the New England market structure.

3 Alternative Settlement Approaches

A basic choice in any energy market is the settlement system. NEPOOL has proposed a single-settlement system. Others have adopted a multi-settlement system. We discuss the issues in both below.

3.1 *Single-settlement system*

In a single-settlement system, the day-ahead bids are used for scheduling, but prices are determined ex post based on real-time dispatch. A single-settlement system consists of the following steps:

- Bids and schedules are submitted day-ahead.
- ISO schedules units for the next day to minimize costs, given the bids, forecasts, operating and transmission constraints, and bilateral schedules.

see Bernard, et al. (1998); Kagel and Levin (1997); and Weiss (1998b). For evidence from the U.K. market, see

- ISO may accept bid/schedule changes up to an hour before real time.
- ISO dispatches units in real time at least cost, given the bids and forecasts for subsequent hours.
- ISO determines real-time spot prices as shadow prices from the actual real-time LP optimization of dispatch.
- Real-time spot prices are used for all settlements to pay generators and charge load.
- Compliance penalties are assessed against those failing to perform as scheduled.

3.2 Multi-settlement system

In a multi-settlement system, the day-ahead bids are used for both scheduling and settling day-ahead transactions. Only deviations from the day-ahead schedule are priced ex post. The steps are as follows:

- Bids and bilateral schedules are submitted day-ahead.
- ISO schedules dispatchable units for the next day to minimize costs, given the bids, bilateral schedules, and forecasts.
- ISO determines the prices associated with the day-ahead schedule as shadow prices obtained from the day-ahead LP optimization.
- The day-ahead prices and scheduled quantities are used in the first settlement.
- ISO may accept bid/schedule changes up to an hour before real time.
- ISO dispatches units in real time at least cost, given the bids, schedules, and forecasts for subsequent hours.
- ISO determines real-time spot prices from the actual dispatch.
- Deviations from day-ahead schedules are settled at the real-time spot prices (second settlement).

A three-settlement system is the same, but with an hour-ahead settlement for deviations from the day-ahead schedules, and then a real-time settlement for deviations from the hour-ahead schedules.

To illustrate how the bidding in a multi-settlement system works consider the energy market with demand-side bidding and unconstrained transmission. One day ahead, suppliers submit supply bids and demanders submit demand bids for energy for each hour and each location. Participants submitting bilateral schedules indicate the amounts to be injected and withdrawn at each location in each hour. From these bids, the ISO constructs the aggregate supply and demand curves, and identifies the market clearing price. Supply bids below the clearing price and demand bids above the clearing price are scheduled. These bids are financially binding.

Green (1996, 1997); Green and Newbery (1992); Wolak and Patrick (1996); and Wolfram (1996).

In the subsequent hour-ahead and/or real-time market, deviations from this schedule are remedied using adjustment bids; that is, incremental or decremental bids that indicate the prices at which the supplier (demonder) would be willing to increase or decrease its injections (withdrawals). These “incs” and “decs,” depending on the design, may be voluntary or mandatory, and in the latter case may be deemed to be the same as the original bids.

The purpose of the incs and decs, however obtained, can be seen in an example. Suppose that a generator fails to deliver. The incs and decs are used to identify the suppliers that are best able to adjust their supplies to balance the market. In this way adjustments are made at least cost. Moreover, deviations from the day ahead schedule are properly priced. If a generator fails to deliver, then other generators will be increased (according to the incs bid), pushing up the spot price. The generator pays a penalty equal to the difference between the spot price and the day-ahead price times the quantity the generator failed to deliver.

3.3 Is a single-settlement system sufficient?

The single-settlement system may appear simpler than multi-settlement systems. First, it involves just a single set of hourly prices. Second, it is closer to the way NEPOOL operated before restructuring. However, this simplicity is deceptive. The difficulty with the single ex post settlement is that much is riding on the ex post prices, since all earlier commitments and transactions are settled at the prices established in real time. After the day-ahead schedule is formed, bidders have an incentive to make adjustments to influence the spot price in a favorable direction. Since the spot price is used for all trades, the incentive for manipulation may be large.⁶ For instance, day-ahead transactions including bilateral transactions may account for 95% of trades, but these are settled at prices that reflect heavily the 5% traded in the real-time market. Bidders can take advantage of short-term inelasticities in the supply schedule to reap excess profits. Knowing how to do this is complex, and can be exploited best by large bidders with sufficient scale to make the efforts worthwhile. The added complexity and risk tends to discourage entry and participation by small bidders whose net revenue might be whipsawed by price volatility in the real-time market.

⁶ Contracts for differences may make the spot price irrelevant for a significant portion of transactions. However, other bilateral transactions may be tied to the spot price. Also the spot price may indirectly influence the terms negotiated in bilateral contracts. Contracts for differences (CFDs) are bilateral hedging contracts. The seller bids supplies into the energy market and the buyer purchases from the energy market. The seller receives the spot price; the buyer pays the spot price. However, the CFD has a strike price known only to the contract participants. If the strike price is above the spot price for the hour, the buyer pays the seller the difference; if the strike price is below the spot price, the seller pays the buyer the difference. The net payment is the contract strike price.

This gaming can be mitigated by financial penalties for failures to perform as scheduled. But then the question is: How to set the penalties? Some flexibility is needed because of uncertainties in demand and supply. Setting the penalties too high leads to inefficient responses to this uncertainty, and setting the penalties too low leads to excessive gaming. The reliance on penalties is highly inefficient and problematic in its workings. It is a carryover from the tight power pool of NEPOOL, and is unworkable on a sustained basis in a competitive market. The whole idea of relying on administered penalties is inefficient, subject to disputes between NEPOOL and ISO New England, and subject to continual pressure (as in Alberta) to seek modifications and exceptions. Compliance is shown to be a problem in Victoria, where a supplier can and does curtail generation by claiming an operating problem, etc.

A multi-settlement system mitigates gaming on two fronts. First, the day-ahead bids are binding financial commitments. The bids and resulting schedules are credible precisely because they are financially binding. Second, bidders are unable to alter the day-ahead prices. These remain fixed for all transactions scheduled in the day-ahead market. Deviations from the day-ahead schedule affect the spot price, but the spot price is used only to price these deviations. Hence, in a multi-settlement system the incentive to manipulate the spot price is not magnified as it is in a single-settlement system.

Penalties for non-performance are not needed in a multi-settlement system, since deviations from the schedule are priced correctly. If a generator fails to deliver as scheduled, then that generator is liable for the spot price for the quantity it was supposed to deliver.

The multi-settlement system reduces risk for the bidders, since the bidders can lock in the day-ahead prices. For the ISO, the multi-settlement system reduces scheduling uncertainty because it discourages schedule changes, and it automatically sets the right penalties for non-performance. The system maintains the flexibility required to respond efficiently to fluctuations in demand and supply.

A difficulty with the multi-settlement system is that it involves multiple prices for energy. One might think that energy at a particular time (and place) should have one price. However, this is not correct. The price should be determined at the time resources are committed. Hence, if there are two commitment points (day-ahead based on forecasts and real-time based on events), then there should be two prices, one a forward price for early commitments and a second that recognizes the effects of contingencies.

Despite the advantages of multi-settlement systems, single-settlement systems can perform adequately for at least a short period of time. The United Kingdom, Victoria, and Alberta provide examples of such systems. However, there is a strong tendency to move away from single-settlement systems. The United Kingdom intends to switch to a multi-settlement system, as does PJM. Both California and New York use or will use multi-settlement systems. We believe that this shift toward multi-settlement systems reflects the significant advantages they offer. The use of single-settlement

systems in some markets is largely a historical artifact. Single-settlement systems are evolutionarily closer to the organization of power pools before competitive wholesale markets were introduced.

The differing incentive effects of the alternative settlement systems are illustrated in the Alberta (single-settlement) and California (multi-settlement) designs.

The design of the California PX may seem awkward at first, and indeed it is awkward in terms of the software required for settlements, since each MWh of energy might be assigned any one of several prices. In the PX's energy market, one clearing price is financially binding for trades completed in the day-ahead forward market, another clearing price is binding in the hour-ahead forward market, and the spot price in real time applies to ancillary services and supplemental energy purchased by the ISO. On the other hand, the advantage of this design is that traders have an incentive to bid seriously in each of the forward markets, since the trades concluded there are financially binding at the clearing price in that market.

Alberta uses the opposite design in which all settlements are made at the final spot price, calculated ex post. That this design produces incentive problems can be seen in the rules required to implement it. Traders were originally prohibited from altering their day-ahead commitments, but then pressures from suppliers led to a compromise in which each trader was allowed a single re-declaration, and lately the argument has been over whether the final time for all declarations should be moved to just two hours before dispatch. These developments reflect all suppliers' preferences to delay commitments until close to the time at which prices for settlement are established, so that uncertainty is reduced, and each supplier's advantage from committing last so that it can take maximal advantage of the likely pattern of prices thereby revealed. The Alberta design also invited a kind of gaming. Importers and exporters are allowed to submit multiple "virtual" declarations. They have used this opportunity to declare several alternatives on a day-ahead basis and then to withdraw all but one shortly before dispatch in order to obtain the best terms. Of course the other traders in Alberta now want the same privilege.

Our opinion is that the difficulties implementing the Alberta design are intrinsic to any design in which transactions are not financially binding at the clearing price in the market in which they are made. Having the day-ahead bids clear at the spot price, rather than the day-ahead price, introduces a basic conflict. One can argue that a sequence of binding forward prices might sacrifice some efficiency in coordinating the day-ahead and real-time markets, as compared to one in which settlements are based only on spot prices, but our view is that this sacrifice is necessary to ensure that bids are serious in the forward markets. If viable forward markets are unnecessary, as perhaps in a purely hydro system, then spot-price settlements are sufficient, but it seems to us that justifications for forward markets also justify binding transactions at the clearing prices in these markets. One must, of course, ensure that the sequence

of forward markets is sufficiently contestable to enable arbitrage that keeps forward prices in line (in expectation) with subsequent spot prices.

Recommendation: *The ISO should switch to a multi-settlement system. We recommend either a two-settlement system with financially binding day-ahead bids, or a three-settlement system with binding bids both day-ahead and hour-ahead. Such an approach provides market incentives for participants to respond efficiently to uncertain demand and supply. Moreover, it mitigates incentives for gaming and reduces bidder uncertainty. Deviations from day-ahead schedules are priced by the market.*

4 Transmission Congestion

A second key issue is how to handle transmission congestion. Ideally, there will be excess transmission capacity and congestion will be important only in exceptional circumstances. In this case, there is just one liquid and competitive electricity market. Energy prices vary by time but not by place. This optimistic view may fit the New England market, where the grid was designed as a single unified market. However, it is worth discussing congestion issues in the event transmission congestion becomes more of an issue after restructuring. Indeed, there is reason to believe that transmission congestion *will* become more of an issue after restructuring. In a market, suppliers will operate to maximize profits. Supplier profits may be enhanced by exploiting transmission congestion to create a local monopoly. For example, a supplier with multiple units may self-schedule them in such a way as to force a unit with a higher bid on line when it otherwise would not be dispatched. Hence, the fact that transmission congestion was not a problem in the past does not mean that it can be ignored in the future when suppliers face market incentives. In any case, it is clear that congestion is likely on the interties that enable imports and exports to adjacent control areas, since there will be strong incentives for arbitrage between the markets in neighboring jurisdictions.

4.1 Uniform uplift

NEPOOL has proposed a uniform uplift charge to cover the costs of alleviating congestion, which is similar to its current practice. Uniform uplift works as follows. When transmission constraints are violated, units are selected out of merit order to solve the security-constrained problem at least cost. The out-of-merit units are paid their bids,⁷ rather than the lower clearing price for energy or the higher clearing price for the marginal cost of adjustments invoked to eliminate congestion. The cost of these extra payments is recovered through a uniform uplift charge that all loads pay on a proportionate basis.⁸

⁷ Provided the market structure and pricing constraints are satisfied; see footnote 3.

⁸ Currently, the security constrained dispatch is not automated. It is done by ISO staff each day.

The uniform uplift approach is simple—all loads pay the same energy price. However, it does not send the correct market signals to demanders *or* suppliers. In a congested system, injections and withdrawals at different points in the network impose different costs or benefits on the system. Efficiency requires that the prices parties face reflect these differences. If one party's additional demand for transfer creates transmission congestion, then it should pay the added costs of adjusting others' generation; if another's additional demand reduces transmission congestion, then it should reap the benefit of the reduced costs of adjustment. Locational pricing is one approach to solve this problem. It is described in the next subsection.

The uniform uplift approach is vulnerable to bid manipulation. A supplier will submit a high bid in situations where it is likely it will be constrained on, and so receive its bid, rather than the clearing price. And a supplier may adjust other bids to increase the chances it will get the constrained-on payment for alleviating congestion. This distorts the energy market, since the same bids are used. The market monitoring rules that limit the situations in which a bidder is paid its bid mitigates the distortion. However, these administrative rules are ad hoc and inflexible. A market based response to the problem is better.

Uniform uplift charges are especially problematic with respect to imports and exports. To the extent that there are price differences between New England and the New York, Quebec, and New Brunswick control areas, there is an incentive for imports or exports.⁹ Uniform uplift may not allow the prices at the interconnections to balance, which typically requires that imports and exports must be rationed administratively using a pro forma tariff, rather than by markets. Administrative procedures may be slow to respond and distort incentives, and thus may undermine the contestability of the New England energy markets. They also create rents that attract strategic behaviors by participants designed to capture them, for example creating congestion to get units with higher bids to run.

Uniform uplift charges suffice provided transmission congestion is not a significant issue. This will be true if uplift charges represent a tiny fraction of the energy cost. However, if the uplift becomes significant, then measures should be taken to adopt a market-based approach to transmission congestion. Since developing and implementing such measures will take considerable time and extensive regulatory approvals, the effort should begin at first sign that the uniform uplift approach may be inadequate. The import/export interfaces deserve continuing close scrutiny.

⁹ The Quebec interface, which will become available for market use in 2000 or 2001, will need to be treated differently, since it is a DC line and therefore does not interact with the rest of the transmission system.

4.2 *Location-based pricing*

In its purest form, real-time congestion pricing of scarce transmission capacity sets a usage charge for each directional link in the system, or equivalently (using Kirchhoff's Laws) an injection charge at each node. The choice between these is often based on practical considerations: there may be many more links than nodes, thereby favoring nodal pricing, but perhaps only a few links are congested recurrently, in which case link pricing is simpler.¹⁰ More frequently, only a few major links or nodes are priced explicitly, and for forward markets it is sufficient to establish injection charges only for nodal hubs or for large zones, or usage charges for major zonal interfaces as in NordPool and California.¹¹ These practices have important implications for the specification of firm transmission rights (FTRs) and price hedges such as transmission congestion contracts (TCCs); for example, secondary markets are illiquid or inactive if the FTRs or TCCs are specified in point-to-point terms rather than zone-to-zone. In principle TCCs are required for every nodal or zonal pair but in practice it suffices to consider only those nominated by traders, and then issue a subset consistent with the system capacity and security constraints. Due to loop flow, a TCC can have a negative value and require the holder to pay rather than receive a usage charge; if this is impractical then the ISO must absorb the cost, whereas the prices of directional links are always nonnegative.

In a competitive market, injection or usage charges are derived from the marginal costs of alleviating congestion, not a tariff or "postage stamp" based on embedded cost. In an optimized pool the usage charge represents the shadow price on transmission capacity, but in decentralized markets it represents the difference at the margin between the cost to the ISO of scheduling an increment (say, of supply in an import zone) and the revenue from a decrement (of supply in an export zone), or the reverse in the case of a demand inc/dec pair. For example, in a two-zone situation the usage charge for the zonal interface is typically the difference in terms of \$/MWh between the most expensive increment in the import zone and the least profitable decrement in the export zone, among those scheduled by the ISO. When the configuration is more complicated the ISO uses an OPF or other optimization program to select the bids that are accepted, taking account of loop flow and security constraints. Congestion pricing in this fashion is based on the principle that the transmission system is an open-access public facility in which (non-discriminatory) charges are imposed only to alleviate congestion on over-demanded interfaces, and

¹⁰ When only a few links have positive prices it is still true that nearly all nodes have nonzero injection charges.

¹¹ In these systems the ISO absorbs the cost of real-time intrazonal balancing via the real-time balancing market.

represent the marginal costs of the re-scheduled resources. In particular, the owners of transmission assets cannot withhold capacity nor affect prices.¹²

A common implementation of locational pricing is locational marginal pricing (LMP). This approach has been adopted by PJM and New York. With LMP, a different price for energy is calculated at each node. These location-specific prices reflect the marginal cost of an additional MWh of energy at each node, taking into account all costs and benefits of the additional energy on the system. In the ISO New England system the LMP would be the location-specific shadow price for energy in the 5-minute dispatch LP. LMP sends the right price signals provided bilateral contracts also pay congestion charges equal to the difference in the LMPs at the two locations. Although locational marginal pricing may appear complex, the prices are easily calculated as part of the dispatch LP, and can easily be made available to participants on a five minute basis via an OASIS bulletin board. PJM's experience with LMP in its first two months, April-May 1998, demonstrates that it can be implemented, even with thousands of nodes, and may be effective in shifting behavior in response to transmission congestion.¹³

Correctly pricing congestion can have substantial benefits in constrained systems. Most importantly, it gives the ISO a market-based means of satisfying security constraints. Participants react to the location-specific prices in ways consistent with economically efficient dispatch subject to these constraints. The alternative of administrative controls is apt to be far less responsive or efficient.

One difficulty with LMP or other location-specific pricing is that it exposes the participants to another source of price uncertainty. Most systems that have adopted location-specific pricing have proposed that tradable fixed transmission rights (FTRs) be issued for some fraction of capacity. These rights include insurance against usage charges in the form of a TCC, and sometimes also scheduling priority as in California. By acquiring FTRs, a participant obtains a financial hedge against the uncertainty of usage charges. An FTR from point A to point B refunds the difference between the LMPs at B and A. Hence, if a unit at A holds an FTR from A to B for 100 MW, then the unit can deliver energy to B at a rate of 100 MW and be guaranteed the price at B. Alternatively, if the 100 MW represents a bilateral transaction, then the parties at A and B are able to negotiate a firm price unaffected by transmission congestion.

Although improving price certainty may be desirable, there is a cost. Namely, the FTRs make the holders immune to the locational prices. If the entire grid capacity is allocated as FTRs, then the ISO loses

¹² An exception is that some owners of transmission assets or grandfathered entitlements, such as municipal utilities, can opt whether to assign their capacity to the ISO for transmission management. If they choose not to do so, then the ISO accepts their schedules without any pricing of congestion.

¹³ Hogan (1998).

its ability to control the system with locational prices. Hence, if FTRs are used, quantities of FTRs should be limited to some fraction of the grid capacity. In this way, those negotiating bilateral transactions can obtain price certainty without sacrificing the security of the grid.

Lastly, we mention a problem with transmission markets based on congestion prices. When usage charges are derived solely from the costs of alleviating congestion, traders can opt to “self-manage” congestion by curtailing their proposed transfers sufficiently to eliminate usage charges. This is unlikely at the level of a small individual trader unless charges are imposed at the level of injection nodes or particular links. But with large zones, market makers conducting exchanges or bilateral contracting that account for large fractions of transmission demand can self-manage in an explicit attempt to capture congestion rents.¹⁴ The California design encourages self-management, and indeed there is no efficiency concern about who captures the rents provided congestion is alleviated one way or another. In contrast, efficiency in an optimized pool requires that all congestion rents are captured via usage charges. This depends on a naïve view of incentives and strategic behavior unless market power is so dispersed that price-taking prevails. More likely, the opportunity to capture congestion rents encourages concerted efforts to capture them.¹⁵

4.3 Incentives for new transmission capacity

The incentives for new transmission capacity will play a critical role in determining the long run efficiency of the wholesale electricity market in New England. There is a strong conflict of interest between suppliers and demanders on this issue, as well as among suppliers.¹⁶ Transmission congestion creates congestion rents. The specifics of the grid and the market rules impact how those rents are distributed among participants. Inevitably some parties will benefit from the status quo, and will lobby hard against changes to the grid. This is especially true of transmission expansion that would increase the contestability of the market from imports. For the electricity market to be efficient, it is essential that such lobbying efforts prove unsuccessful. At a minimum, new capacity should be added whenever its cost can be covered by a reduction in future congestion costs; more generally, capacity expansions need to take account of the inherent externalities that pervade transmission systems. Indeed, if there is a bias in these

¹⁴ This is not necessarily easy to do, since there is a significant free-rider problem engendered by each exchange’s preference that others bear the greater share of the burden in curtailing their aggregate transmission demands. The game is repeated daily, however, so implicit collusion is potentially feasible.

¹⁵ Theoretical models as well as experimental and empirical results indicate that energy traders capture some portion of congestion rents. See footnote 4 and Weiss (1998a).

¹⁶ The transmission cost sharing and planning mechanism that is being developed may resolve some of these conflicts, since it divides power between demanders and suppliers.

decisions it should be toward building excessive transmission capacity, since this capacity is essential in making the markets contestable.

Resolving this issue is beyond the scope of this report. It is largely a governance issue,¹⁷ and is not affected directly by the market rules. (One exception is the debate over uniform uplift charges or locational pricing of congestion. Locational pricing has the advantage that it makes the congestion costs transparent and therefore is likely to stimulate demand for new transmission capacity in the places where it is most needed, and to encourage siting of new generation capacity in locations enabling unconstrained exports.) We mention the issue because of its importance. It would be pointless to make a Herculean effort in fine-tuning the market rules if the goal of long run efficiency is undermined by poor governance rules.

Recommendation: *We doubt that uniform uplift will resolve congestion problems adequately. The ISO should begin investigating alternatives. Location-based pricing will increase short-run efficiency, and provide improved incentives for generation and transmission expansion. It may suffice in the short-term to impose congestion charges only on the import/export interfaces of the New England system.*

5 Interaction between Energy and Reserve Markets

In considering the interaction between the energy and the reserve markets, it is important to keep in mind that incentive effects are not eliminated by one market design or another; rather, the form in which they are expressed depends on the specific features of the market structure. The advantage of a superior design derives from the extent to which it enables traders to express accurately the economic considerations important to them. Gaming strategies are inherent in any design that requires traders to manipulate their bids in order to take account of factors that the bid format does not allow them to express directly.

The bid format is a key factor. For example, if the market is organized to provide hourly schedules and prices, then this tends to serve the interests of demanders for whom the time of power delivery is important, and suppliers with flexibility (for example, storage hydro), whereas it tends to ignore the considerations of suppliers from thermal sources, who are mainly concerned with obtaining operating schedules over consecutive hours sufficient to recover the fixed costs of startup and who are unconcerned about timing per se. Schemes have been devised that allow demanders to bid on a time-of-day basis while suppliers bid for operating runs of various durations; prices can then be stated equivalently in terms of hourly prices for demanders and duration prices for suppliers. Similarly, for ancillary services it is usually important to distinguish between availability payments for reserving capacity and payments for delivered

¹⁷ See, for example, Barker, Tenenbaum, and Woolf (1997).

energy when called by the ISO. Schemes have also been devised to allow bids in terms of priorities or adjustments, such as demands that are curtailable above a specified real-time price. We bypass these more elaborate schemes here in order to focus on the basic problem of clearing an hourly market for firm energy, either forward or spot.

A pool tries to eliminate inefficiencies by a centralized optimization based on submitted cost and engineering data, some of which is monitored for accuracy. The program allocates quantities subject to system constraints, but it also obtains shadow prices used for settlements. In principle, a dual formulation could be implemented as a single market with explicit prices determined by simultaneous clearing of the markets for each of the main ingredients, such as energy, transmission, and ancillary services. Several designs have been proposed for conducting these markets simultaneously, and at least one has received some experimental testing. In practice, however, these markets are usually conducted in a sequence reflecting the fact that demands for transmission are derived from energy transactions, and the supply is fixed. Similarly, the ISO's demand for ancillary services is nearly proportional to the demand for energy, since most system operators maintain reserves on that basis, and the supply consists mostly of residual generation capacity after accounting for the main energy transactions. Thus, the typical structure is a cascade in which the initial market is for energy, followed by a transmission market in which energy flows are adjusted to keep within the transfer capacity, then a market for ancillary services such as spinning and non-spinning reserves (for which some transfer capacity was previously set aside). These forward markets on a day-ahead (and perhaps hour-ahead) basis are followed by a real-time market in which the ISO draws on supplementary offers to maintain system balancing on a short time scale, and when these are insufficient or expensive, calls on the ancillary services held in reserve.

5.1 Do the rules promote intermarket efficiency? In particular, can energy and reserve markets be effectively unbundled?

Ancillary services are especially sensitive to the bid format. Using spinning reserve as the example, it is clear that suppliers must be paid for capacity availability as well as energy generation. On this basis one might surmise that suppliers should bid both components, and then these two-part bids should be weighed together by the ISO in bid evaluation, but this causes problems. The initial problem is that the independent system operator must evaluate such two-part bids by giving some weight (interpreted as the probability or duration that spinning units will be called to produce) to the energy bid. But as in most multi-part bidding schemes, this is fraught with gaming problems; for example, a bidder who thinks that a call is less probable than the weight used by the ISO prefers to exaggerate the capacity bid and shrink the energy bid, and the opposite if a call is more probable. Thus the merit order of energy bids reveals less

about actual costs of generation than expectations about the likelihood that spinning reserves will be activated.

These incentive problems are alleviated when different procedures are used for bid evaluation and settlements, as in California.¹⁸ In the simplest scheme bids for reserves are accepted solely on the basis of the offered capacity price, and then settlements for energy generation are based on the system real-time energy price rather than the offered energy price.¹⁹ That is, the offered energy price is interpreted only as a reserve price below which the supplier prefers not to be called. Thus, it provides a merit order for calling generation without distorting incentives. This scheme separates the competitive process into two parts corresponding to the two parts of the bid, one for capacity availability, and another for priority in being called to generate. The result is economic efficiency in both the energy and reserve markets, provided the markets are competitive. Bidders have the incentive to bid their marginal cost in the energy market, and their opportunity cost in the reserve market. Like California, the proposed market rules for New England also have this desirable structure. We return to this topic in Section 8.

6 Revelation of Bidding Information

In each of the electricity markets, an important question is: What bidding information should be revealed to the bidders, and when and how should it be revealed? Since the issues are similar across markets, we begin with a general discussion, but frame it in terms of the energy market to be concrete. We describe two extreme approaches, and then discuss a middle ground that is more apt to be appropriate for ISO New England.

For any piece of information received or produced by the ISO, there are several options.

1. The ISO can reveal publicly the information (public information).
2. The ISO can reveal the information to all the bidders, but not the public.
3. The ISO can report the information to the specific bidder (bidder-specific information).
4. The ISO can report the information to no one (secret information).

We do not give option 2 a name, since it is an option that can be dispensed with immediately. Any information that is revealed to all the bidders should be made public. The reason is that if it is useful information to the bidders it is useful information to a *potential* bidder. Since the ISO cannot know the set

¹⁸ Wilson (1998b).

¹⁹ One qualification to this statement is that bids that would not be least cost for any real-time price are screened out before ordering the capacity bids in merit order.

of potential bidders, the information should be made public. However, the decision among options 1, 3, and 4 is less obvious.

At one extreme is a fully transparent process: reveal all bidding information to the public. This is the approach adopted by the FCC in its highly successful spectrum auctions. Before the auction begins the FCC posts the set of eligible bidders, the extent of each bidder's eligibility, the bidder's identity, and the bidder's application form, which includes ownership and other financial information. During the auction, after each round of bidding, the FCC immediately posts all the bids for the round, the bidder that made each bid, and any changes in eligibility. This approach has three main advantages. First, it gives participants (and potential participants) the maximum amount of information. The bidders then can use this information in preparing subsequent bids. The information reduces the bidders' uncertainty, which facilitates price discovery and improves efficiency. Moreover, it may increase auction revenues, since with less uncertainty the bidders can bid more aggressively, without fear of falling prey to the winner's curse.²⁰ Second, it simplifies implementation. The information simply is posted on the Internet. The FCC need not worry about maintaining or delivering bidder-specific information. Nor does the FCC need to be concerned with establishing secure methods of preserving secret or bidder-specific information. Third, it means that the process is fully transparent. This permits the bidders and any other interested party to check that the auction is being conducted in compliance with the stated rules. If problems are discovered, they can be fixed quickly, before any serious damage is done.

The difficulty with a fully transparent process, which prove critical in electricity markets, is that information is sometimes a two-edged sword. It can be used to facilitate explicit or implicit collusion, as well as promote efficiency. Information about the bidder identity associated with each bid is especially vulnerable to implicit collusive use. For example, a group of bidders can establish a collusive supply schedule, and then punish defections to the schedule. If bidder identities are known, then the punishment can be directed against the defector, by retaliating in particular hours or locations, so as to harm the defector the most. Alternatively, a small subset of bidders may be party to a tacit agreement. For example, only the three largest bidders may have a implicit collusive understanding. In this case, to enforce the collusive arrangement, it is important for the colluding bidders to know the bidder identities, so that deviations can be detected, and then punished.

²⁰ The winner's curse is the tendency for naïve auction winners to lose money, because they fail to take account of the information contained in winning a competitive auction. To avoid the winner's curse, smart bidders shade their bids. The amount of shading depends in part on the amount of uncertainty the bidders face. See Milgrom and Weber (1982).

In a few of the FCC spectrum auctions, some bidders did take advantage of the fully transparent process to send messages to their rivals, telling them on which licenses to bid and which to avoid.²¹ These bidding strategies may have helped these bidders coordinate a division of the licenses, and enforce the proposed division by directed punishments. The FCC's experience stands as a lesson that profit-motivated bidders in high-stake auctions take advantage of opportunities that are permitted by the rules.

At the other extreme is a policy of complete secrecy. The ISO makes no disclosure of any information, aside from what is absolutely necessary—each bidder would only be told its settlement information. This approach would mitigate tacit collusion to the greatest extent. However, it exposes the bidders to the greatest uncertainties, and this may introduce inefficiencies.

A middle ground is probably best. First, the secrecy of individual bids is essential for competition in this kind of market. The market is repeated daily and a few participants constitute the majority of supply. Such a setting is ripe for abuse if the parties are given the informational means. System-wide results should be public information: prices, total generation, total reserves, etc.²² This information is either needed by bidders for planning or can be inferred from settlement information. Hence, it should be made public. The next step would be to make the aggregate bid schedules public. Bidders surely would like to have this information in preparing bids for the next day. It represents an indication of what would be the consequences of changing the quantity bid. However, the information is not essential for competition. A supplier, whose bid was rejected yesterday, knows that it needs to improve its bid tomorrow (assuming tomorrow is like today). Knowing the price elasticity of supply (or demand) is not essential to the analysis. Unless a strong argument can be made that knowing the price elasticity improves efficiency, it would seem prudent to keep the aggregate schedules secret and only reveal prices and aggregate quantities publicly.

One argument in favor of revealing more information is that it may level the playing field. Each of the two largest bidders has knowledge of nearly one-half of the bids, and a combined bid knowledge of 65% of the bids. In contrast, the other bidders know only a tiny fraction of the bids. One solution to this problem would be to introduce an information policy for co-owners on units for which they do not have bidding authority. For example, one could delay the release of bid information to non-lead co-owners until the end of the month. This would give the co-owners the information needed for oversight of the lead owner, but prevent the use of the information for daily strategic gain.²³ An alternative would be to make the aggregate bid schedules public information.

²¹ Cramton and Schwartz (1998a,b).

²² The revelation of imports/exports, especially those in short notice, needs more careful analysis.

²³ The feasibility of this option would depend on overcoming legal and enforcement difficulties.

Recommendation: *The risk of implicit collusion is sufficiently large to outweigh the efficiency gains of disclosing information beyond market prices and total quantities.*

7 Energy Market

In the next three sections we discuss market-specific issues raised by the proposed rules. We begin with the energy market.

The proposed energy market is a single-settlement system, which is cleared ex post. Except for the case of out-of-merit dispatch, all suppliers receive the same energy price—the real-time spot price. This reliance on the spot price for all trades in the ISO energy market creates a strong incentive for suppliers to influence the spot price through actions taken after the day-ahead schedule is announced. If the spot market is thin, participants may be whipsawed by large price variations. Below we discuss the calculation of the energy clearing price and several other important issues in the energy market.

7.1 Calculation of the energy clearing price

Because so much is riding on it, the calculation of the energy clearing price (ECP) in a single-settlement system is a frequent source of debate. In one view, the real-time spot price should be the marginal cost of producing an extra MWh of energy given the state of the system, whatever the costs of reaching that state. The alternative argument is that the ECP should be the bid of the marginal unit that is dispatched—that is, the highest accepted bid for the time period—taking account of ramping constraints then in effect. Practically, these alternative methods of calculating the energy clearing price (ECP) are implemented as follows:

1. *Shadow price method.* The ECP is the shadow price on the energy balance constraint from the 5-minute dispatch LP.
2. *Marginal unit method.* The ECP is the most expensive MW from all dispatchable units, including any ramp-constrained units, that was actually dispatched to meet energy demand.

The central issues are how intertemporal constraints are handled, and how gaming is to be prevented. If there are no intertemporal constraints, then the two methods yield the same outcome, both in pricing and dispatch. This would be the case if all units could respond flexibly in a 5-minute interval, so that ramping constraints were nonbinding. However, when ramping constraints are binding, the two approaches differ.

Assume that participants bid their costs and that demand is known perfectly in advance, so that re-dispatch to meet contingencies is unnecessary. Then the “correct” ECP is the shadow price on the energy balance constraint in each 5-minute interval from the scheduling LP over the *full cycle* (24 hours) with all

ramping constraints included.²⁴ Then the shadow price during a 5-minute peak includes the costs of previously ramping up and then later ramping down expensive units to meet the spike in demand. For practical reasons (e.g., uncertainty in demand), the proposed shadow price method uses the shadow price from the 5-minute dispatch LP. Hence, the LP is taking the initial state as given, and looking forward it takes demand predictions as accurate, ignoring the fact that units are in a particular state now as a result of prior optimization and that subsequent contingencies might require further alterations in the planned dispatch. The shadow prices are thus different from the ideal. However, we conjecture that over a daily cycle that is approximately repeated each day, the proposed shadow price approach simply displaces some payments from one hour to another, and is unbiased on average in that the payments summed over the full cycle will be the same as in the full-cycle LP.²⁵ In some hours the price will be too high (e.g., price in peak periods do not account for ramping costs incurred earlier and later) and in others too low, but these discrepancies from the ideal balance out over the daily cycle. This is the primary justification for the shadow price method.

It should be understood that the distortions in the shadow price method, even if unbiased on average, are relevant to participants. In particular, how the distortions impact a participant may depend on how flexible the participant's plants are. Different unbiased approaches may result in more or less variation in spot prices. Those participants with more flexible plants may be better able to take advantage of a more variable spot price, especially high prices in peak periods. A simulation study would be required to say much more about how different ECP calculations would impact participants.

Recently, a "Compromise Proposal" has been discussed.²⁶ The compromise method expands the set of units eligible to set the ECP, but imposes two screens to mitigate risks of gaming. The method allows flexible units—those that are not ramp-rate constrained due to changes in their energy bid prices—to set the ECP. A unit is deemed flexible if its manual response rate (MRR) is not too low. A unit is deemed ramp-rate constrained due to changes in its energy bid price if (1) it is ramping down and its price schedule increased from the prior hour, or (2) it is ramping up and its price schedule decreases in the next hour. The flexibility screen is argued to prevent a bidder from strategically setting a low MRR. The

²⁴ We are assuming (in the simplistic world without uncertainty) that the scheduling program breaks the 24-hour day into 5-minute intervals, and then solves the dynamic programming problem with all 5-minute ramping constraints included. This problem produces a shadow price on energy for each five minute interval that reflects intertemporal constraints. Of course, it is not practical to solve such a problem, so the unit commitment programs use 1 hour intervals. Also, demand and supply uncertainties make the day-ahead schedule only a rough approximation of what is required.

²⁵ A true optimization using LP under uncertainty (Dantzig, 1976) that includes contingent plans would provide fully correct pricing. However, it is presently seen as infeasible computationally.

²⁶ "Compromise Proposal for Determining the Real-Time Marginal Price of Energy," June 1, 1998.

pricing screen is argued to prevent a bidder from strategically changing its bid schedule to affect the clearing price when ramp-rate constrained. Under the compromise proposal, the ECP is set at the higher of (a) the shadow price from the 5-minute dispatch LP and (b) the highest cost unit dispatched, including ramp-constrained flexible units that pass the pricing screen.

This compromise adds complexity to the rules for price determination, and it may bias the ECP in an upward direction. The two screens, while intended to mitigate gaming, are ad hoc, and have the undesirable consequence of imposing two artificial constraints on the bidders. To the extent that these constraints distort bidding behavior, efficiency is compromised. For example, bidders may limit changes in their bid schedules across hours, despite changes in their marginal costs, so as to qualify to set the clearing price.

The debate over the clearing price calculation highlights the weakness of a one-settlement system: too much is riding on the spot price. To the extent that parties transact through the ISO's residual energy market, the spot price will determine all settlements, and thus participants may have strong incentives to distort their behavior in order to manipulate the spot price. Because the largest bidders have significant market shares, we anticipate that manipulation of the spot price will be an important issue. A full analysis of the various pricing methods would take into account how the pricing rule impacts the parties' incentives to game the system. Such an analysis is beyond the scope of this report.

One consistent view of the scheduling-dispatch problem is as a day-ahead scheduling LP with uncertainty that is resolved at dispatch. Hourly shadow prices on the energy balance equation in the scheduling LP represent correct energy prices at the day-ahead point. Then the 5-minute dispatch LP (which is a contingent sub-problem of the larger scheduling LP), *taking the state as given*, produces the correct energy prices for deviations from the day-ahead schedule. Hence, a multi-settlement system in which prices from the scheduling LP are used for day-ahead transactions, and then prices from the real-time dispatch are used for deviations from the day-ahead schedule will produce more accurate incentives. Notice that the shadow price and marginal unit approaches nearly coincide in the multi-settlement system: the day-ahead shadow price is the cost of the marginal unit at the time of commitment. To a large extent, the debate over shadow price methods vs. marginal unit methods is resolved in a multi-settlement system. The difficulty with the single-settlement approach is that one is asking the 5-minute dispatch LP to produce prices it is not designed to produce—prices that reflect the economic costs of committing a unit on a day-ahead basis to hourly service whose costs are dependent on the state of the system when contingencies must be addressed.

Recommendation: *ISO New England should move to a multi-settlement system. Although it could not be implemented by December 1, its implementation is helped by the fact that the neighboring power markets in PJM and NY have adopted a multi-settlement system.*

7.2 Is the lack of demand-side bidding a serious flaw in the energy market?

Demand-side bidding is required for full efficiency. Short-run efficiency requires demand-side bidding, and long-run efficiency requires incentives for investments in cost-effective price-sensitive demand reduction technologies. Technological advances in the next few years will increase the elasticity of short-run energy demand by enabling faster responses to price variations. Indeed, most of the efficiency gains in the long run are likely to come on the demand side rather than the supply side. Demand-side bidding is essential to obtain these potential efficiency gains. Demand-side bidding creates incentives for investing in power management technologies that economize on energy consumption in peak periods. Without demand-side bidding these innovations will be stifled.

Market power mitigation is inherently more difficult when demand is treated as inelastic. A key result from experimental studies is that demand-side bidding is a powerful instrument in mitigating market power.²⁷ Investments in power management technologies to increase demand elasticity would limit supplier market power. This is likely to be as effective in reducing supplier market power as investments in new generating capacity. Of course, this is possible only if demand-side bidding is introduced.

Finally, ISO operations are impaired when it must rely on predictions for its optimization procedures, and cannot use demand-side management options such as curtailments. Although curtailable loads are allowed under the proposed rules, demand-side bidding is a more flexible market mechanism to account for buyers' sensitivity to price.

One explanation offered for the absence of demand-side bidding is that allowing it would be overly complex. This explanation does not stand up to the fact that many electricity markets have successfully implemented demand-side bidding at reasonable cost. There is no reason not to allow demand-side bidding. Its absence is an artifact of the era of regulation, which focused on the supply side, taking the demand as given.

Recommendation: *Demand-side bidding should be introduced as soon as possible. The ISO should develop the rules and other steps needed for implementation, taking advantage of the experience of other markets, such as California.*

²⁷ Bakerman, et al. (1997) and Weiss (1998b).

7.3 Will the energy market be adversely impacted by short-notice external energy transactions in that purchases and sales without corresponding self-schedule changes are allowed?

Short-notice transactions together with ex post settlement will introduce opportunities for severe gaming. For example, a large bidder can bid in such a way that the supply schedule in the day-ahead market is steep at the clearing price. Then selling a large quantity to New York as a short-notice external transaction has a large price effect. Indeed, it gives other large bidders an incentive for similar external transactions. Since the market is repeated on a daily basis, and prices are fairly predictable, large bidders will likely learn the advantage of submitting steep supply schedules around the clearing price, and then driving the price high on all units using appropriate external transactions. Since these short-notice transactions can amount to 10% of the market or more, the price effect can be very large.

On the other hand, such strategies would be mitigated by short-notice imports from New York, as traders take advantage of arbitrage opportunities. In this case, the short-notice transactions may work to make the New England market contestable. The critical elements in evaluating whether short-notice transactions will help or hurt competition are the rules and charges that govern use of the transmission capacity for imports and exports. FERC's open access order 888 mandates that this critical transmission capacity should be used efficiently to arbitrage price differences across markets. However, there still may be impediments to effective arbitrage. For example, if a large bidder in the New England market schedules imports from New York until the interconnection is constrained, and then makes a last minute short-notice export to New York, units in New York may have insufficient time to take advantage of the arbitrage opportunity. Such a strategy might work especially well if the neighboring market has predominately inflexible generation or internal transmission constraints (and corresponding usage charges) that are binding.

The procedures for accepting short notice transactions can mitigate the problems they may cause.²⁸ However, it would be impossible to develop workable administrative procedures that effectively differentiated between short-notice transactions motivated by desirable arbitrage and those motivated by strategic gaming.

Recommendation: *Adopting a multi-settlement system would eliminate the inherent gaming problems that short-notice transactions create. Under a single-settlement system, restrictions would need to be put in place to mitigate gaming. These restrictions would limit flexibility, creating new inefficiencies. Congestion pricing of import/export interfaces would improve the efficiency of their use.*

²⁸ "Internal Guide for Short Notice Transactions," ISO New England.

7.4 Will a residual energy market establish a reliable spot market price for electricity?

The two requirements for reliable prices are competition and liquidity. These two features usually go hand in hand. Both competition and liquidity may be an issue in New England. First, the two largest participants supply over half the market, ignoring imports. These are likely sufficient shares to influence prices at times of peak demand. The absence of demand-side bidding will accentuate any market power problems as described above. Second, liquidity may be a problem under the proposed design. There may be a strong incentive for participants to use bilateral transactions and to self-schedule. Bilateral transactions avoid risks associated with reliance on the spot prices, which are inherently volatile. This risk will be especially high if competition in the spot market is weak. Bilateral transactions also have the advantage to the parties that self-schedules can be made ignoring any transmission costs they impose on the system. Since transmission costs are borne on a proportionate basis via uplift charges, the parties to a bilateral transaction pay only a small fraction of the costs they impose on the system. In contrast, if the parties submitted bids into the ISO energy market, then their bids might be rejected, and out of merit bids accepted in their place due to transmission congestion. These factors provide incentives for bidders to avoid the ISO energy market by self-scheduling bilateral transactions.

On the other hand, bilateral transactions without self-scheduling do have one important advantage. The supplier can satisfy the demand via the ISO energy market if it is efficient to do so. For example, the parties can use a contract for differences to avoid the risk of an unreliable spot price, and still take advantage of the efficiency gains from not self-scheduling. In fact, an unreliable spot price increases the incentive to do so, since the gains from market arbitrage increase with the market's volatility. This incentive provides a countervailing force that prevents the energy market from collapsing altogether.

Still, on balance, the ISO's share of energy transactions is likely to fall to a tiny fraction. This has been the experience in other single-settlement systems, such as Victoria, where only about 5% of the system's energy is traded in the ISO's market. As the share falls, the spot price may become unstable, and the ability of the ISO to control the system using market mechanisms may be compromised. In the worst case, the ISO is forced to use administrative procedures that are fraught with inefficiencies. Alternatively, the growth in the market for bilateral contracts or other commodity markets based on a market-clearing exchange may make the ISO market either unstable or irrelevant, as is the case of the government auctions for sulfur-dioxide allowances.²⁹ Instabilities in a tiny ISO energy market might be avoided through arbitrage from a secondary market. However, the prospect of transmission congestion could prevent this optimistic scenario from materializing. The self-scheduled transactions may prove infeasible,

²⁹ Joskow, Schmalensee, and Bailey (1996).

and an ISO relying on uniform-uplift congestion pricing lacks the market mechanism to correct the problem.

The absence of location-based congestion charges makes the decline of the ISO energy market likely. Private markets do not need to worry about transmission in uniform-uplift systems like New England, and the ISO handles all ancillary services, etc. We expect a dramatic expansion of private contracts and commodity markets, and a corresponding shrinkage of the role of the ISO itself, with resulting thinness and volatility of its real-time spot prices. However, these prices will become inconsequential if most trades are bilateral. The ISO likely would become a minimal-ISO engaged mainly in real-time balancing and ancillary services. These activities will become difficult as the percentage of self-scheduled transactions increases.

Recommendation: *The spot price is likely to be volatile, and possibly unreliable, in a single-settlement system without congestion pricing of transmission.*

7.5 Does the lack of explicit start-up and no-load costs in the bidding mean that efficiency will be sacrificed?

A peculiarity of some optimized pools is payment to suppliers for capacity in addition to energy, based on so-called multi-part bids that include components for both fixed costs and incremental energy costs, with compensating charges to demanders for “uplift.” These are not payments for capacity reserved for ancillary services but rather for planned generation. This holdover from the era of regulation is unique to the electricity industry, which is the only one that does not expect suppliers to cover fixed costs, such as capital and maintenance, from the market price of its output. Although a long-run equilibrium in the industry implies prices in peak periods adequate to cover the costs of capacity idle in other periods, the motive for these payments is apparently the short-run concern that market-clearing prices for energy will be determined by incremental generation costs that will be insufficient to recover the costs of capital and O&M. Such an outcome is mainly a consequence of reliance in optimized pools on shadow prices that reflect only purported incremental costs, based on a parallel optimization of unit commitments that takes account of start-up costs, ramping constraints, and minimum generation levels, as well as the uncertainty of demand and the imputed value of lost load.³⁰ Without elaborating fully here, we are skeptical of any such payment scheme that is not tied to explicit reservation of capacity, such as for ancillary services, because we see it as an open invitation for manipulation. Designs such as those in California,

³⁰ It is also a consequence of relying entirely on supply-side management, taking demand as fixed and inelastic. At the very least comparable payments should be provided to demanders who accept curtailable or lower-priority service. Demand-side measures can reduce the probability and imputed value of lost load, and thereby the reliance on peaking capacity that is idle much of the year.

Scandinavia, and Australia (and recently proposed for the revision of the U.K. system³¹) dispense with these payments by clearing the market for energy entirely on the basis of prices offered for delivered energy, leaving scheduling decisions to suppliers. It might indeed be that prices in California will reflect only incremental costs that are insufficient to recover the O&M costs of installed units, but if so then that signals excess capacity that in the long run should be mothballed or decommissioned. In New England the markets for operable and installed capability may prevent this from happening, as efficiency would dictate.

Recommendation: *We believe that pure energy bids are all that is required for efficiency. The inclusion of start-up and no-load components, while sensible in a centrally planned system, are prone to gaming in a decentralized bidding environment. Bidders should be able to schedule their units reliably through pure energy bids.*

7.6 Is single-round bidding sufficient for economic efficiency?

In energy markets there is a basic distinction between static and iterative market processes. In a static design for a pooled market each trader provides a single bid, usually in the form of a demand or supply function, with or without a separate capacity bid or a minimum revenue requirement, and perhaps in the form of a portfolio bid for multiple generation sources that is only later converted into unit schedules. The static character lies in the fact that the initial market clearing is also the final one. The theory underlying a static design is the Walrasian theory of markets, in which the market finds a price that equates stated demands and supplies. The mode of competition lies in each trader's selection of the bid function it submits—which requires substantial guesswork since others' bids are unknown when the submission is made.

If the bids are purely for hourly energy then a static design can cause problems for suppliers with fixed costs and ramping constraints because the revenue may be insufficient to cover total costs. Designs of this sort therefore provide approximate remedies: the UK provides capacity payments, Spain allows suppliers to specify a minimum revenue requirement, and New England allows minimum run times. Without elaborating details here, our view is that these auxiliary provisions engender as many gaming problems as they solve, and in the case of capacity payments based on an assumed value of lost load, are inherently arbitrary.

An iterative market process works quite differently, and reflects the Marshallian theory of markets. As in an auction with repeated bidding, it is those traders whose bids are at the margin who contend to get their bids accepted, and in each round they can base their bids on the tentative results from previous

³¹ "A Common Model ...", Offer, London, May 1998.

rounds. For example, suppose that as usual a supplier's bid is submitted as a series of steps at successively higher prices. In this case an "extra-marginal" supplier, one with a step above the market clearing price, realizes that by reducing its price for that step it can be more competitive in the next round—thereby ejecting an infra-marginal bidder who in the next round becomes extra-marginal and therefore must itself improve its offered price. Thus, Marshallian competition works by inducing competition among those bidders whose steps are actually near the margin, in contrast with Walrasian competition in which the price offered for each step must be based on a conjecture about the competitive situation in the event that step is at the margin.

Iterative processes require procedural "activity" rules to ensure serious bidding throughout (and thus reliable price discovery) and to ensure speedy convergence, but they have the advantage of avoiding ad hoc measures to assure bidders' fixed costs are covered.³² In a day-ahead auction the key feature is that an iterative process enables "self-scheduling" in the sense that each supplier can adapt its offers in successive rounds to the observed pattern of hourly prices. With good information about the prices it can obtain in each hour, a supplier with steam plants can itself decide on which units to schedule, their start times, and their run lengths. Similarly, a supplier with storage hydro sources can better tailor its releases to take advantage of the observed prices in peak periods. In the California PX this enables pure-energy portfolio bidding: only after the energy market clears do the portfolio bidders need to report to the independent system operator their unit schedules that provide the energy they sold. Instead of the detailed operating data required by the UK's static pool to run its centralized optimization program, California's decentralized design assigns authority to the suppliers to schedule their own units to meet the commitments contracted in the energy market.

These considerations are not unique to the operation of markets organized as exchanges with an hourly market clearing price that applies uniformly to all trades. Most markets for bilateral trades allow a dynamic process in which bid and ask prices are posted continually, and any posted offer can be accepted at its offered price at any time. As in an exchange using an iterative market clearing process, traders can monitor the posted prices and the prices of completed transactions to obtain good information about the prevailing pattern of prices. And because the contracts are bilateral, each party can set its own schedule to

³² The activity rules for the California PX are adapted from the FCC's auctions of spectrum licenses, which have been notably successful and are now used worldwide. The PX rules were tested in laboratory experiments at Caltech with good results, but they will not be implemented in the PX until late 1998, so there is presently no factual evidence on their performance in practice.

fulfill the bargain. There are also designs for bilateral markets in which all contracts are tentative until the market clears, and then the same hourly prices apply to all completed transactions.³³

Although some of the advantages of iterative bidding are clear, there are special features of the energy market that may make iterative bidding unnecessary or impractical. An energy auction must be conducted quickly. It would be impractical for the bidding to last more than an hour or two, which allows only a few rounds of bidding. The repeated nature of the energy auction also makes iterative bidding less essential. If inefficiencies are discovered one day, they can be corrected in the bidding for the next day. The daily repetition of the energy auction provides much of the benefits of learning during an iterative process. Inefficiencies may occur, especially in response to unexpected demand or supply shocks, but they may be corrected the next day.

The California PX has been working fine without iterative bidding since it began on March 31, 1998. Although the California market was designed with iterative bidding, this format will not be introduced until late 1998, due in part to delays in software implementation. California participants view the absence of iterative bidding as of secondary importance. Their view is that the multi-settlement design gives them ample opportunity to resolve inefficiencies. If a unit goes unscheduled in one hour but is scheduled in the adjacent hours, then the unit can bid the missing hour in the hour-ahead market. If it still remains unscheduled, it can bid the missing hour in the real-time balancing market. Hence, the 3-settlement system gives the bidders additional opportunities to make adjustments (hour-ahead and spot markets), which mitigates the need for iterative bidding. New England's proposed single-settlement system gives the bidders fewer opportunities to correct problems, so a stronger case for iterative bidding can be made in New England.

Recommendation: *If a multi-settlement system is adopted, bidders will have opportunities to correct scheduling infeasibilities and inefficiencies. In this case, iterative bidding in the energy market is probably unnecessary.*

8 Ancillary Services

The argument is occasionally made that an energy exchange might as well augment each demand bid by the required proportion of ancillary services, or at least spinning reserve—just as is typically done for transmission losses.³⁴ This argument recognizes that on the demand side spinning reserve is a necessary

³³ This design has been studied experimentally in the University of Arizona laboratory, but we have not seen a practical implementation.

³⁴ Most systems assign to suppliers an approximate cost of losses, without attempting an exact calculation. In California, for instance, a “generation meter multiplier” is assigned to each node and updated continually to account partially for losses, and the residual is absorbed by the ISO.

complement to planned energy deliveries. It is mistaken, however, because on the supply side energy and spin are largely substitutes, not complements. Moreover, technologies differ considerably in their characteristics for spinning reserve; for example, storage hydro sources and fast-start turbines are not subject to the ramping constraints and no-load costs of steam plants, but on the other hand, thermal plants can provide spinning reserve by operating below capacity. It is better therefore to establish a separate market for spinning reserves (and curtailable loads) along with other ancillary services so that these differing characteristics can be reflected in bids.

Unlike the energy market, where payments are made for deliveries, payments for ancillary services are based on scheduled reserves of capacity, which may or may not be called for energy generation. As a result, sufficient monitoring and penalties for nonperformance are critical to assuring that the scheduled ancillary services are actually available when called. These issues are addressed in rules 13 and 14. In this section, we will assume that the monitoring and enforcement is sufficient to make the bids in these markets credible.

8.1 Pricing of Ancillary Services

The pricing of ancillary services specified in the NEPOOL Market Rules and Procedures, Section 6, is the most unusual feature of ISO New England's proposed pricing structure.³⁵ We have serious reservations that are described below.

It will suffice here to consider a thermal unit reduced from its high operating limit (HOL) to provide TMSR.³⁶ Unlike a hydro unit that can submit a positive bid for reserving capacity for TMSR, a thermal unit's bid is deemed to be zero. Its payment per hour for each MW of capacity reserved for TMSR is essentially

$$2 \times \text{Max}\{ 0, \text{Estimated Spot Price for Energy} - \text{Submitted Bid Price for Energy} \},$$

and in addition it receives the actual spot price for energy produced if and when it is called for TMSR generation in real time. This reservation payment has several peculiar features:

- The estimated spot price is based on an optimization that ignores the requirements for ancillary services, so it presumes that the cost of the marginal generator is the same with and without set-asides for ancillary services.
- The factor 2 represents the fact that nearly the same payment is calculated twice.

³⁵ These rules are amplified in the document "Reserve Market Example: Ten Minute Spinning Reserve."

³⁶ The reduction is the MW that can be ramped in ten minutes, based on the manual response rate, and sustained for at least thirty minutes or longer as required.

1. A *de facto* bid is calculated as the lost opportunity cost, interpreted as the unit's foregone profit from running below its HOL. This profit is calculated as the difference (if positive) between the estimated spot price and the unit's bid price.
2. This cost is then computed again, but using the bid-in cost of the "next least expensive MW dispatched" in place of the estimated spot price, and again credited to the unit as an estimated "production cost change."

The *de facto* bid (1) might be an appropriate compensation for reserving capacity for TMSR. Indeed, if no formula payment were provided in the rules, then this is approximately what the unit's owner would want to bid for reserving capacity for TMSR, since it represents the profit foregone from not selling generation in the energy market. The *de facto* bid is actually only an upper bound on this foregone profit, however, since in fact the unit may also earn additional profit if it is called for TMSR generation, for which it is compensated at the spot price.

On the other hand, the "production cost change" (2) is a misnomer, since it bears no relation to any actual change in production costs. It might be considered a subsidy for standing ready to ramp up if called, but otherwise we are at a loss to see its motivation.

The net effect of this double counting is that the TMSR selling price includes (besides the actual bid if the unit is hydro) the real-time price twice, once as a component of the lost opportunity payment, and again as part of the lost opportunity clearing price.

We see no economic justification for the inclusion of both the lost opportunity cost and the so-called production cost change. If we are correct in our reading of the procedural rules, then the net effect is that for reserving capacity for TMSR a thermal unit is deemed by the formula to have bid twice an estimate of what it foregoes by not selling generation directly in the energy market.

Since we are unfamiliar with the origins of this payment formula we must rely on some guesswork at this stage. The basic problem seems to arise from the fact that a thermal unit submits only a bid for energy generation. Because the ISO then selects some units for TMSR it must provide compensation for forcing the unit to operate below its HOL or desired dispatch point. This compensation is bounded above by the difference between an appropriate estimate of the spot price and the unit's bid price—if in fact it is not subsequently called for generation. The true compensation required is this upper bound less the profit from generation when actually called. If this interpretation is correct then it appears that the bonus or subsidy provided TMSR units for standing ready is the sum of the production cost change and any profits subsequently earned when called to generate.

The compensation for reserving capacity for TMSR that is used by other ISOs is considerably different—and settlements are much simpler because the payments rely on a uniform price rather than

unit-specific payments. For example, California pays the spot price for energy generated just as does ISO New England, but for the capacity reservation it pays only the highest bid among those selected, and rather than a *de facto* bid the unit's actual bid is used for the selection. In particular, this enables the bidder to make its own estimate of both the foregone profit from direct energy sales, less the expected profit in real time from called generation. It should also be mentioned that most systems do not experience a shortage of resources for TMSR since in any case those units whose capacity is not fully sold in the energy market are available, and their opportunity cost is by definition zero.

Double counting can potentially produce significant price distortions. It appears optimal to underbid a unit's marginal cost of generation near the HOL so as to increase both the opportunity cost and the production change cost components. This risks being excluded from the TMSR selection, and lowers the estimated spot price, but these risks can be more than compensated by the double counting.

Why hydro can bid but thermal cannot is unexplained, but one can surmise that hydro is being allowed the option to opt out of the selection for TMSR dispatch (by pricing itself out of the market) because of operating or intertemporal constraints, or total-energy constraints, but it is peculiar because most systems think of storage hydro as ideal for TMSR. Note too that thermal units sufficiently below HOL receive *no* opportunity cost payment.

Recommendation: *The TMSR pricing should be greatly simplified. At the very least, the double-counting should be eliminated. A preferred approach would be to let all capable resources bid for TMSR and let the bid reflect any opportunity cost in this or other energy markets.*

8.2 Unbundling Energy and Reserves

The unbundling of energy and reserves need not be a source of inefficiency. Efficiency is attained by separating the markets as follows: The reserve bids are used to determine which capacity reservations are accepted. Reserve bids are accepted in the merit order determined by their capacity prices, up to the total required by the ISO. All accepted offers are paid the market clearing capacity reserve price (the highest accepted bid). Then the energy bids determine which units provide energy. Energy bids are accepted in merit order, and bidders are paid the real-time spot energy price for any energy delivered.

The efficiency of such a system may seem counterintuitive, since it makes no effort to optimize jointly the energy and reserve markets. However, it is precisely this joint optimization—selecting energy and reserve resources to minimize total system costs—that creates the incentives for bidders to distort their bids, undermining efficiency. With efficient pricing, the only joint optimizing feature is to screen out bid pairs that are never cost minimizing.

This approach is extended easily to the practical setting with multiple reserve options. These options—TMSR, TMNSR, and TMOR—should be thought of as a cascade of options the ISO draws on to cap the spot energy price. Bids rejected for one service, say spinning reserve, can be carried over to compete for another service, such as non-spinning or operating reserve. In addition, the ISO should set the demands for these services based on the submitted schedules. For instance, because TMSR is superior to, and can substitute for, TMNSR, the ISO should purchase TMSR rather than TMNSR up to the point that TMSR becomes more expensive to meet the need for non-spinning reserve. In general, each superior-quality service should be allowed to substitute for inferior qualities whenever it is cheaper. In this way, the clearing price in each market will reflect the quality of service: the clearing price for TMSR will be at least as great as that for TMNSR, which will be at least as great as that for TMOR.

Recommendation: *Demands for ancillary services should be set in response to the submitted schedules. In particular, the ISO should select the percentages of spin, non-spin, and operating reserves optimally to take advantage of the offered bids. In this way prices decrease in the sequence of reserves, reflecting the quality of the service.*

9 Capacity Markets

In principle, the installed capability and the operable capability markets are unnecessary in a competitive electricity market. Under competition, capacity is determined over the long term by the market in response to price expectations. If expectations are correct, then sufficient capacity is built so that the market prices just cover all costs including a risk-adjusted return on capital investments. If expectations are incorrect, then prices will be high, prompting additional investment in capacity, or low, prompting curtailed investment in capacity.³⁷

The capacity markets are a holdover from the regulated setting, when capacity decisions were not made in response to price expectations. In the transition to a competitive market, the capacity markets may serve a useful role in coordinating investments in capacity. However, once competitive electricity markets are established in New England, it would be appropriate for the capacity markets to terminate.

One possible advantage of the operable capability market is that it may reduce price peaks by facilitating the coordination of maintenance schedules. However, it does nothing to prevent strategic quantity reduction by suppliers, since a supplier can make “operable” capacity inoperable by submitting a sufficiently high energy bid. Unless there are constraints on energy bids, the operable capacity market

³⁷ The fact that proposals totaling 25,000 MW of new generation were received before the capacity markets were formed or functioning suggests that generators will respond to high price expectations with new generation.

does little to prevent price peaks caused by suppliers withholding quantity. However, it may discourage suppliers from scheduling maintenance at peak times or the same times.

Despite any advantages the capacity markets may have in transition, it may be best to eliminate them before the December 1st start date. The reason is that once these markets are established there will be parties that benefit from them at the expense of other participants. For example, incumbent suppliers may want capacity requirements to continue so that they receive payments for obsolete plants that they would otherwise decommission, to be replaced by new more efficient plants. These parties will lobby for the status quo, and organizational inertia may mean that the capacity markets last well beyond any usefulness in transition. The costs from having the capacity markets are increased complexity, and distortions in investment decisions. Capacity costs should be recovered through prices in the energy and reserve markets, and not as part of an artificial market created by administrative regulations.

Since the installed capability market is a monthly market and the stakes are high, it may make sense to allow iterative bidding. An iterative market has many advantages—most importantly a more reliable process of price determination. Bidders are able to respond to tentative price information in the prior rounds. Uncertainty is reduced, which encourages a more efficient outcome.

An iterative auction could be conducted in a few hours on the last day of each month. Internet-based auction software, which is available now, could be customized to this auction market. This would allow bidders to participate from their corporate headquarters (or any other location worldwide).

To implement iterative bidding two additional changes would be needed. First, demand-side bidding would be introduced. And second, the ex post settlement would be replaced by a two-settlement system. The auction on the last day of the month would determine the market clearing price for installed capability based on the bids and offers in an iterative process. To the extent that there are deviations ex post, then the deviations are priced according to incs and decs specified by the suppliers. These changes would be needed to have a meaningful price discovery process with iterative bidding.

Recommendation: *The ISO should consider eliminating the installed capability market and the operable capability market. Incentives for sufficient capacity are provided by the energy and reserve markets. If the installed capability market is retained, consider switching to iterative bidding.*

10 Conclusion

We believe that the wholesale electricity market in New England can begin on December 1, 1998. However, improvements are needed for long-run success. We have identified four major recommendations:

- Switch to a multi-settlement system.

- Introduce demand-side bidding.
- Adopt location-based transmission congestion pricing, especially for the import/export interfaces.
- Fix the pricing of the ten minute spinning reserves.

Of these, only the final (and least important) recommendation could be implemented by the December 1st start date. We do not view this as a fatal problem, provided that by the start date the ISO and NEPOOL reach agreement in principle on these basic concepts and a tentative timetable for implementation. We believe that if the markets open without any sense of what improvements will be made or when they will be made, then it will be much more difficult to adopt and implement needed improvements. An evolutionary “wait and see” approach would be too slow, and likely would result in damage to the markets that is difficult to correct.

Our examination of the architecture of wholesale electricity markets in New England presumes that the ingredients for effective competition are present. It is important to emphasize further that market architecture is distinctly secondary in importance to market structure, in the sense of competitiveness or contestability. Monopoly power in generation, or local monopolies due to transmission constraints, can impair efficiency regardless of the market design implemented. Oligopolies are inherently more damaging to the public interest in wholesale electricity markets because their daily interaction offers ample opportunities for punishment strategies to police collusive arrangements, whether explicit or implicit. Thus, structural solutions to the market power of dominant incumbents often are important. How much competition is enough is a difficult empirical question. The ISO’s plans for market monitoring are an essential element to resolving this question in New England.

In the same way, procedural rules are less important than architecture: no amount of fiddling with procedural rules can overcome major deficiencies in the links among the energy, transmission, and ancillary services markets. There is therefore a natural priority in the design process that starts with ensuring a competitive market structure, proceeds to the selection of the main market forums, and then concludes with the detailed issues of governance and procedures. Some procedural rules, of course, must be designed to mitigate market power and prevent collusion; for example, it is usual to maintain the secrecy of submitted bids to thwart efforts by a collusive coalition to punish deviants.

An aspect omitted here is the role of transaction costs. This consideration affects all three stages of the design process. Procedural rules must obviously be designed to avoid unnecessary transaction costs, but it is well to realize too that a complex array of decentralized markets imposes burdens on traders, who may well prefer a simpler structure that avoids managing a complex portfolio of contracts, bids, and schedules. A simple design can also promote competition by bringing all traders together in a few markets

with standardized contracts, bid formats, and trading procedures. The virtues of simplicity should be kept in mind as the market rules are refined. Simple rules encourage participation; whereas, complex rules stand as a barrier to entry.

References

- Ausubel, Lawrence M. and Peter Cramton (1998), "Demand Reduction and Inefficiency in Multi-Unit Auctions," Working Paper, University of Maryland.
- Bakerman, Steven R., Michael J. Denton, Stephen J. Rassenti and Vernon L. Smith (1997), "Market Power In A Deregulated Electrical Industry: An Experimental Study," Working Paper, University of Arizona.
- Bakerman, Steven R., Stephen J. Rassenti and Vernon L. Smith (1997), "Efficiency and Income Shares in High Demand Energy Networks: Who Receives The Congestion Rents When a Line is Constrained?" Working Paper, University of Arizona.
- Barker, James, Bernard Tenenbaum, and Fiona Woolf (1997), "Governance and Regulation of Power Pools and System Operators: An International Comparison," Working Paper, The World Bank.
- Bernard, John C., Richard Zimmerman, William Schulze, Robert Thomas, Timothy Mount, and Richard Schuler (1998), "Alternative Auction Institutions for Purchasing Electric Power: An Experimental Examination," Working Paper, Cornell University.
- Cramton, Peter and Jesse Schwartz (1998a), "Collusive Bidding in the FCC Spectrum Auctions," Working Paper, University of Maryland.
- Cramton, Peter and Jesse Schwartz (1998b), "Collusive Bidding: Lessons from the FCC Spectrum Auctions," Working Paper, University of Maryland.
- Dantzig, George B. (1976), *Linear Programming*, Princeton, NJ: Princeton University Press.
- Green, Richard (1996), "Increasing Competition in the British Electricity Spot Market," *Journal of Industrial Economics*, 44, 205-216.
- Green, Richard (1997), "The Electricity Contract Market in England and Wales," Working Paper, Cambridge University.
- Green, Richard J. and David M. Newbery (1992), "Competition in the British Electricity Spot Market," *Journal of Political Economy*, 100, 929-953.
- Hogan, William W. (1998), "Getting the Prices Right in PJM: Analysis and Summary, April-May," Working Paper, Harvard University.
- Joskow, Paul L., Richard Schmalensee, and Elizabeth M. Bailey (1996), "Auction Design and the Market for Sulfur Dioxide Emissions," Working Paper, MIT.
- Kagel, John H. and Dan Levin (1997), "Independent Private Value Multi-Unit Demand Auctions: An Experiment Comparing Uniform Price and Dynamic Vickrey Auctions," Working Paper, University of Pittsburgh.
- Milgrom, Paul and Robert J. Weber (1982), "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50, 1089-1122.
- Weiss, Jürgen (1998a), "Congestion Rents and Oligopolistic Competition in Electricity Networks: An Experimental Investigation," Working Paper, Harvard University.
- Weiss, Jürgen (1998b), "Market Power Issues in the Restructuring of the Electricity Industry: An Experimental Investigation," Working Paper, Harvard University.
- Wilson, Robert (1998a), "Priority Pricing of Ancillary Services in Wholesale Electricity Markets," Working Paper, Stanford University.

- Wilson, Robert (1998b), "Efficiency Considerations in Designing Electricity Markets," Report to the Competition Bureau of Industry Canada, March 31.
- Wolak, Frank A. and Robert H. Patrick (1996), "The Impact of Market Rules and Market Structure on the Price Determination Process in the England and Wales Electricity Market," Working Paper, Stanford University.
- Wolfram, Catherine D. (1996), "Strategic Bidding in a Multi-Unit Auction: An Empirical Analysis of Bids to Supply Electricity in England and Wales," Working Paper, Harvard University.