

REGULATORY EVOLUTION, MARKET DESIGN AND UNIT COMMITMENT

Richard P. O’Neill, Udi Helman and Paul M. Sotkiewicz
Federal Energy Regulatory Commission

Michael H. Rothkopf
Rutgers University

William R. Stewart, Jr.
The College of William and Mary

Abstract: In the context of competitive wholesale electricity markets, the unit commitment problem has shifted from a firm level optimization problem to a market level problem. In some centralized market designs it is used to ensure reliability and determine day-ahead market prices. This paper reviews the recent history of short-term electricity markets in the United States to evaluate the experience with alternative market designs and the implications for unit commitment modeling. It presents principles for the design of the next generation of unit commitment-based markets.

1. INTRODUCTION

Competitive wholesale electricity markets are now operational in several major U.S. markets, confirming the analysis and recommendations of prescient economists, electrical engineers and others over the past two decades.¹ Since generation comprises approximately 75 percent of all electricity costs, competition in generation promises large efficiency gains and cost savings to consumers. The unit commitment problem, the traditional method by which regulated utilities and power pools conducted internal

¹ Seminal contributions on competitive wholesale electricity markets include [1,2]. As of January 1, 2000, regional markets with centralized wholesale electricity exchanges are operational in California, the Pennsylvania-New Jersey-Maryland (PJM) interconnection, New England, and New York.

scheduling of generation to meet demand at least cost over a multi-hour to multi-day time frame, is now embedded, in various ways, in competitive markets. Potentially, unit commitment models will be used by different market participants and institutions: individual firms, centralized auctioneers, decentralized aggregators of generation schedules, and transmission system operators. This environment presents a new set of modeling requirements and market design challenges.² The market level unit commitment problem is typically much larger in scale than the firm level problem. Speed and accuracy are important if the solution is being used by an auctioneer to determine market prices. Evaluating the characteristics of the solution, such as the presence of duality gaps (implying a lack of market clearing prices) and alternative optima, becomes of direct financial interest to market participants.

The new electricity markets, and hence new applications of the unit commitment problem, are being developed within an evolving regulatory context. Indeed, an important driver of market designs is the guidance given by the regulator. The Federal Energy Regulatory Commission (henceforth “the Commission”) initiated regulatory reform of transmission in 1996, with the objective of encouraging competitive regional electricity markets that promote economic efficiency without compromising system reliability. The regulatory approach, embodied in a series of orders described below, has been to provide an open market architecture within which alternative market designs are being implemented, evaluated, and changed when necessary. Research into the unit commitment problem has largely been reactive to the new regulatory environment and the emerging issues in market design. A more proactive approach is needed. Among the issues that need consideration and research are the choices between simultaneous and sequential optimization of several energy and ancillary service products, alternative bidding rules for different products, different mechanisms for congestion pricing, and inter-regional coordination. In addition, unit commitment modeling now has to confront the issue of economic incentives in various market settings, which requires a more extensive familiarity with economics and game theory.

In response to the regulatory evolution it has set in motion, the regulator also needs to adapt institutionally and develop its technical capabilities. This is imperative because the Commission is taking an oversight role in market design decisions across the United States. Several wholesale markets operate centralized unit commitment auction markets (e.g., PJM, New England and New York), the market design of which is the focus of this

² The literature on market design in electricity markets is extensive; for a survey, see the articles in [3].

chapter. These markets also allow bilateral trading and types of self-scheduling. Following approval of the basic design, the Commission provides oversight for a flood of subsequent adjustments and refinements in the search for well functioning markets. The underlying unit commitment model is often either an implicit or explicit matter in these market rule decisions.

The objective of this chapter is to describe regulatory evolution and the market design challenges for unit commitment modeling. The chapter focuses on day-ahead markets, but much of the discussion is also applicable to real-time markets. Section 2 of the chapter describes the key regulatory developments and the design and recent experience of the major regional wholesale electricity markets. Section 3 focuses on principles that should guide the design of day-ahead energy and ancillary service markets. Finally, Section 4 offers conclusions.

2. REGULATORY EVOLUTION AND THE ORGANIZATION OF ELECTRICITY MARKETS

The recent history of electricity regulatory reform in the United States began when the Commission issued Orders 888 and 889 in 1996 [4,5]. These orders required an open access transmission regime, based on non-discriminatory transmission rates and transparent posting of available transmission capacity (ATC). Order 888 also included fairly broad organizational principles for an Independent System Operator (ISO), an institution which separates ownership from control of the grid and can perform market functions. Between 1997 and 2000, ISOs and PXs were formed in California and in the three tight power pools of the eastern United States. These ISOs established day-ahead and real-time markets for energy, ancillary services and transmission (in California, the day-ahead energy market is conducted by several separate scheduling coordinators, including the California Power Exchange). In the Midwest, an ISO has also been established, but is not yet operational. Other regions of the country have been less successful or unwilling to centralize grid operations and electricity trading remains bilateral, with a vertically integrated utility performing the balancing and reliability functions.

By early 1999, a certain amount of inertia was evident in the development of wholesale markets. Electricity traders were expressing dissatisfaction with the traditional methods of transmission grid management still employed in large parts of the United States. Specifically, there was substantial concern about frequent curtailments of transactions, justified on the basis of reliability

but often questioned by parties to the transactions.³

On December 15, 1999, the Commission took a step toward clarifying the appropriate transmission access and market institutions with Order 2000, which requires the formation of Regional Transmission Organizations (RTOs) [6]. The order establishes in the RTO many of the features that had emerged in the ISO markets as well as additional characteristics and functions that address unresolved issues in both ISO and non-ISO electricity markets. The RTO is required to serve a region of sufficient scope and configuration to provide for a reliable, efficient electricity market. With respect to the unit commitment problem, some of the important features of the RTO are that it must have exclusive authority for maintaining short term reliability, act as provider of last resort of ancillary services, address parallel path flows, provide real time energy balancing, and ensure development of market mechanisms for congestion management.

In many ways, the functions assigned to the RTO are based on the principles of market design embodied in the better functioning ISOs that have emerged (RTOs will, in fact, subsume existing ISOs). Section 2.1 surveys some of these ISO market design lessons; Section 2.2 discusses further the objectives of Order 2000 and some market design issues.

Before considering the details of open access and market design, an important question is: Why should the regulator remain involved in the design and oversight of the emerging competitive markets? Recent experience has made clear that, in near term, the Commission has a continuing role for at least three reasons:

1. There is the ongoing development of open access itself, including the provision of short-term reliability services linked to transmission and future expansion of the grid. There are public good aspects to reliability and transmission expansion.⁴
2. Where not currently available, there must be an efficient pricing mechanism for transmission congestion (i.e., pricing of the externalities created by parallel path flows).
3. There must be mitigation of market power in markets for electricity and ancillary services and in provision of transmission capacity.

³ Such curtailments are supposed to follow the North American Electricity Reliability Council's (NERC) Transmission Loading Relief (TLR) procedures, which provide criteria for the management of congested transmission facilities.

⁴A public good is a good that is non-rivalrous and non-excludable. In the case of reliability, a load's or generator's consumption of reliability in no way prevents others from enjoying the same reliability, and if all of the loads and generators are interconnected on the same system, they cannot be prevented from enjoying the benefits of reliability.

Market power is the ability of firms to raise prices above competitive levels. Firms can exercise market power in electricity markets because of both structural factors (e.g., firm concentration or transmission constraints) and opportunities offered by the market design (for a survey, see [7,8]).

The Commission's policy heretofore has been to approve implementation of the markets (through more liberal standards for granting market-based rates) while at the same time evaluating the markets' operational experience and providing guidance on market designs as a means to promote efficiency and competition. In addition, a significant amount of the responsibility for day-to-day monitoring and mitigation of market power has shifted to the ISOs and the future RTOs. The regulatory approach, then, seeks to balance the current reality—some firms can exercise a degree of market power generally or under certain system conditions—with the expectation that entry of new firms and more efficient market designs will substantially mitigate future market power.⁵

2.1 EXPERIENCE WITH ISO AND BILATERAL MARKETS

Order 888 outlined principles for but did not require a particular structure for competitive wholesale energy markets. Two broad types of market structures have developed. The first type consists of markets with ISOs, which may or may not include one or more scheduling coordinators or power exchanges (PXs).⁶ The ISO markets with PXs take the form of either a centralized ISO/PX or a decentralized ISO and PX(s). The centralized ISO/PX markets use unit commitment models for creating the day-ahead schedule (which incorporates bilateral transactions and self-schedules) while the decentralized ISO markets require self-commitment by the PXs. The second type of market has no ISO; rather, the transmission system and much of the generation continues to be operated by vertically integrated utilities. Power is traded bilaterally.

This section will focus on the various market designs and performance of the existing ISO markets, as well as offer some conclusions and

⁵ There is a large literature on market power due to both structural and market design characteristics of electricity markets. For analysis of regional energy and ancillary service markets in the United States, see [8-12].

⁶ Scheduling coordinators or power exchanges (PXs) are power trading operations functionally separate from the ISO. These terms will be used interchangeably.

recommendations on what designs will work best. It will also include a brief review of the performance of the bilateral, non-ISO markets.

Market Functions and Design of ISO markets. Competitive wholesale electricity markets can be complex, with multiple interdependent products sold on different time frames and differentially priced at different locations. Existing ISO markets can be characterized by the (1) number and types of different products (energy, ancillary services, capacity), (2) bidding and scheduling process, (3) relationship of temporal (forward and real-time) markets, (4) market clearing and settlement rules, and (5) type of congestion management and transmission rights. Each of the existing ISOs has also established market power monitoring and mitigation, but this function will not be examined here.

The ISO carries out the basic function of assessing the feasibility of proposed generation schedules. The ISO also serves as the buyer, through contracted rates and bid based auctions, of reliability services, including short-term ancillary services (voltage support, operating reserves and automatic generation control) and possibly also longer term capacity products. The PX facilitates and conducts a forward auction market for energy (electric power). PX functions can either be carried out by the ISO itself, as in New York, New England, and PJM, or by one or more separate, unaffiliated PXs, as in California.

With the exception of the California ISO, the ISOs run a unit commitment model to determine which units will be scheduled to provide energy and ancillary services during the following day.

In California, the ISO and PX (which is one of several scheduling coordinators) are separate, a decision intended to keep the transmission system operators (who may have been affiliated with an incumbent utility) functionally distinct from the market. In the California PX market, generation owners self-commit their units through scheduling coordinators. This market structure has experienced certain disadvantages. One problem is that the scheduling coordinators' submissions can be physically infeasible. The ISO must then engage in a time consuming iterative process with the scheduling coordinators to resolve the infeasibilities. However, the PX auction algorithm is such that self-committed units are often asked to start-up and stop, disregarding minimum run and down times, with potentially adverse results.⁷ Another problem is that the separation of the ISO and PX raises the transaction costs for market participants. In addition, because the ISO's decisions about congestion and purchases of ancillary services cannot be

⁷Conversations with California ISO staff confirm this problem.

remedied by the PX, market participants may bid strategically into the ISO's congestion market to ensure that profitable transactions are not curtailed.

Bidding and Scheduling. The ISO markets began, largely, with only supply-side bidding for energy and certain ancillary services. While demand-side bidding for energy is allowed in some markets (currently the California ISO and PX and New York, but planned for the other ISOs), the energy demand function is largely inelastic—that is, not price responsive—due to both technical limitations and historic rate designs. As more price responsiveness is introduced into demand bid functions (through installation of metering equipment and technological advances in distributed generation and information technology), there should be a reduction in both price volatility and the potential for exercise of market power, particularly during peak hours.

The structure of bids is another market design issue that has attracted attention. Bids in current ISO energy markets vary in the number of cost components and required technical parameters, such as ramp rates, high and low operating limits, and so on. In the so-called “one-part” incremental energy bid, the bidder must factor its start-up, no load and other costs into its day-ahead energy bid for each megawatt-hour (MWh) offered.⁸ Even so, generators face the risk that they may not cover all of their costs in the auction. One-part bids require generation owners to internalize this risk in some fashion, which in turn increases their costs (use of the real-time market to make adjustments can eliminate part of this risk).⁹ One-part bids also result in some inefficiency if these are the only costs the dispatcher can consider in unit commitment.

In a three-part bid, the start-up and no-load costs can be separated out, allowing generators to bid actual operating costs more precisely and allowing for a more efficient unit commitment. In PJM and New York, generators are guaranteed to at least recover all of their bid costs if they are committed to run.¹⁰ This mechanism eliminates the uncertainty of whether a generator will

⁸ Currently, the California PX requires one-part energy bids without technical parameters, such as minimum run times, ramp rates, and so on. This could be called a pure one-part bid. New England currently requires one-part bids with technical parameters.

⁹ At its worst, the need to internalize risk in the one-part bidding system could lead to a greater incentive to internalize via bilateral contract or merger to avoid higher transactions costs. High market concentrations lead to market power concerns. Further, not being allowed to bid marginal cost is an easy defense to an inquiry on market power abuse.

¹⁰Generators will receive an “uplift” payment to recover their costs only if the revenues they receive from the energy and ancillary services markets are less than their total bid costs.

be committed and dispatched only to lose money, and it allows for a more efficient dispatch.

Market Clearing and Settlement Systems. Market clearing rules and settlement systems are the procedures that determine quantities produced and consumed, who pays, and who gets paid. As discussed above, ISOs typically operate multiple markets, including energy, several types of ancillary services, and transmission products. There are two basic ways to clear these different markets, sequentially or simultaneously, with variations on each method. In general, sequential auction markets clear each product separately in a sequence, even though several of the products may be alternative (substitute) uses of the same generator. Variations of this approach were adopted initially in California and in the interim New England market. In contrast, a simultaneous auction, adopted in PJM and New York, clears the relevant markets at the same time, minimizing the joint bid cost of providing energy and ancillary services. This method explicitly takes into account that some products are substitutes.

One of the clearest market design lessons from ISOs is that sequential market clearing without product and/or quantity substitution is economically inefficient and offers opportunities for strategic behavior. In California, energy, regulation, 10-minute spinning reserves, 30-minute non-spinning reserves, and replacement reserves are cleared in the order given.¹¹ It has sometimes been the case that the price for ancillary services with lower production costs exceeds the price of ancillary services (and energy) with higher production costs (for example, providing spinning reserve requires using some fuel, whereas providing non-spinning reserve requires simply being on stand-by). The reason for some of these “price inversions” is that generators can strategically bid high prices in the last markets knowing that there will not be much capacity remaining after the other markets clear. Hence, the ISO must take these high price bids. To combat this design flaw, the California ISO has instituted a pre-processing algorithm, called the Rational Buyer Protocol, which will allow it to substitute higher quality services for lower quality services if and only if it reduces its ancillary service procurement cost. New England also initially conducted sequential market clearing and experienced problems similar to California; temporary solutions

¹¹ In California, regulation refers to automatic generation control, but is defined in terms of whether generation output is increased (regulation up) or decreased (regulation down). Spinning reserves is reserve capacity available in a specified time period from a generator synchronized with the grid. Non-spinning reserves is reserve capacity available in a specified time period from a generator not synchronized with the grid. Replacement reserves are reserves that can be available within 60 minutes.

have included rolling over of bids between substitute services, as in California, as well as price caps.

Experience with simultaneous market clearing is limited. The simultaneous market clearing method will largely, but not entirely, avoid the price inversions seen in the California ISO and New England markets because the software used to clear the markets automatically clears all remaining arbitrage opportunities. In addition, generators which bid strategically, as in California, would be far less likely to be selected to provide ancillary services at the higher price due to the greater substitution possibilities in the simultaneous market clearing. While more complex computationally, simultaneous market-clearing appears to be emerging as the better system from an efficiency standpoint.

The settlement systems can be characterized as either single settlement or multi-settlement based on the number of temporal markets the ISO runs. In California there are three temporal markets: day-ahead, hour-ahead, and real-time with a financial settlement for each. New York and PJM run two temporal markets: day-ahead and real-time. New England currently has only real-time markets (bids are due day-ahead but financial settlement only takes place at real-time prices), but is scheduled to implement day-ahead markets with financial settlement. A clear market design lesson is that single-settlement systems, which require generators to submit bids and stand-by day-ahead while awaiting financial settlement at real-time prices, create problems in scheduling and often require additional rules to constrain generator incentives to change their bids. The multi-settlement system has been adopted by all the ISOs in recognition of the value of the forward market as a financial hedge for real-time conditions. Also, the forward market should facilitate demand-side responses by giving demand that has bid to reduce load more time to react to price signals.

Congestion Management and Pricing. In the ISO context, there are two general ways to manage congestion: locational pricing and non-locational pricing. Locational pricing can be sub-divided into approaches defined by the level of aggregation used to calculate the price. In the typical “nodal pricing” method, an energy price is calculated at each generation and load bus (node). The transmission congestion price between any two busses is the difference in energy prices at the busses.

At higher levels of aggregation, the busses in the system operated by an ISO can be gathered into one or more congestion zones. Zones are intended to be indicative of the historical pattern of congestion in the system on the presumption that congestion will take place between the zones with little or no congestion within a zone. The price of congestion between zones is the difference in energy prices between the zones. If congestion occurs within a

zone, the costs of managing it (typically through generator re-dispatch) are shared by all market participants inside the zone using a system of subsidies.¹² This intra-zonal congestion management method could be considered a type of non-locational pricing.

The zonal approach has been adopted in California, which currently has two (soon to be three) “congestion zones” but is experiencing congestion that should trigger new zones (also, each import point effectively creates a new zone). The California ISO manages inter-zonal congestion through adjustment bids submitted by generation and load. These bids indicate the price at which the market participant is willing to be ramped up or down in order to alleviate congested lines. If this fails to relieve the congested lines, then the ISO must call on generators with cost-based contracts to relieve congestion.¹³

California manages intra-zonal congestion by redispatch (which incorporates the transmission constraints into the original, transmission unconstrained dispatch) with the resulting costs averaged over load in the zone. Persistent intra-zonal congestion indicates that the zones are not properly defined; in addition, the averaging of congestion costs within the zone is inefficient, since the congestion costs are also paid by participants not causing the congestion.¹⁴

In the long run, zonal pricing as practiced in California can lead to price signals that distort decisions on siting new generation and transmission assets.¹⁵ Neither maintaining fixed zones in the face of intra-zonal

¹² Several ISOs have attempted to operate as single zones (PJM, New England), but have subsequently made the transition to locational pricing. If the single zone system has consistent congestion between sub-regions (that is, should be at least two zones), this can create opportunities for generators to leave the spot market and use bilateral contracts to take advantage of the system price. This was the experience in PJM before it implemented a nodal system; the ISO was required to adopt administrative measures to curtail the bilateral transactions.

¹³ These are called Reliability-Must-Run (RMR) contracts. RMR contracts are intended to ensure that the ISO has sufficient generation capacity to meet various system contingencies, such as congestion relief and voltage support.

¹⁴ The California ISO can create a new congestion management zone if the cost to alleviate congestion over the previous 12 months exceeds 5 percent of the approximate annual revenue requirement of the transmission operators. In order to be considered an active congestion zone, the markets on either side of the congested interface must be “workably competitive” for significant portions of the year.

¹⁵ For example, the Commission has rejected a California ISO proposal (Tariff Amendment No. 19, filed June 23, 1999) that new generators upgrade transmission capacity to alleviate intra-zonal congestion which might arise from their entry on the grounds that it could create further barriers to entry and market distortions. A similar New England proposal was also rejected.

congestion nor continuous re-zoning are efficient methods of congestion management.

Zonal market design in California has been instituted in part under the rationale that it lowers market power. Both in theory and practice this assumption has been proved wrong. Market power cannot be reduced by the declarations of large zones. If this were so, there would be no market power problem. Transmission constraints and generation costs determine the size of the market, not the declaration of zones. The California ISO rules recognize this by providing for dispatch orders and out-of-market payments to generators in the same zone separated by constraints.

In contrast, despite opposition from some generators and marketers, nodal pricing has been adopted in New York and PJM and approved for New England (these systems actually use nodal prices for generators and zonal averages for loads). Nodal pricing eliminates the problem of properly defining zones, and the need to average the costs of any intra-zonal congestion. In the short run, load receives the proper price signals about how much to consume, and the long-run decisions can be made much more easily. Even though loads pay a zonal price, the nodal price information remains available for decision making. For financial markets, nodal prices for a region can be aggregated into fewer “hub” prices, which are weighted averages of the underlying nodal prices. For example, PJM has two hub prices.

Transmission Rights. Transmission rights have traditionally been used to reserve access to the transmission system and to ensure that energy transactions would be curtailed only in extreme circumstances. These rights were physical rights—the right to physically transmit a specific amount of power over the system for the access charge paid. With the advent of congestion pricing (whether zonal or nodal), most ISOs have provided both physical rights and financial rights that can be used as a hedge against congestion costs (the stochastic nature and potentially high cost of congestion makes financial hedging necessary).¹⁶ In all the markets with locational congestion pricing, payment of congestion prices is essentially a physical right to transmit between nodes or zones (although not a right that is bought in advance). On the other hand, financial rights are typically purely financial mechanisms that provide revenues but confer no physical priority. They can be traded on a secondary market.

For example, in New York and PJM, financial transmission rights give the holder the right to collect congestion rents between a designated point of

¹⁶ Transmission rights can take the form of either options or obligations.

injection and point of withdrawal, so that if a transaction incurs congestion costs, those costs would be offset by the revenues from the financial right. Auctions for these rights are typically held regularly. The California ISO has implemented a similar type of zone-to-zone right, but which also confers some physical priority.¹⁷

Performance of ISO markets. As discussed above, none of the ISO markets has reached a stable point in terms of market design; some are undertaking major market re-designs while others are in the process of implementing major components of their market design. There is a convergence in market design in many areas: all the ISOs have implemented either sequential auctions with substitutions or simultaneous auctions for energy and ancillary services; most ISOs have established multi-settlement systems or will shortly; most ISOs offer some form of financial transmission right; in the East coast ISO markets, nodal pricing is used for generation or is planned for future implementation.

Given these ongoing changes, the preliminary performance of the markets varies by product and time period. In transmission, the ISOs have recorded few curtailments. However, there has been some concern that the number of bilateral transactions has decreased in nodal congestion management systems (because point-to-point congestion may be difficult to hedge with the available transmission rights). The energy markets seem to be functioning fairly well, although prices under certain system conditions reflect varying levels of market power [8-13]. Entry of generation, transmission capacity expansion and demand-side bidding should lower prices and lessen volatility.

The ancillary service markets have been more problematic. Reserve markets in particular have experienced price spikes and price inversions, reflecting the greater vulnerability of these markets to market power and to market design flaws that exacerbate strategic behavior [9,12]. Temporary price and bid caps and more permanent market re-designs should help solve some of these problems. Other general market problems include limitations in software implementation and technical capabilities (such as using telephone rather than electronic communications for dispatch), and conflicts that emerge when system operators depend on rules of thumb to dispatch the system rather than the outcomes of the auction. In general, however, many market design or implementation problems are amenable to satisfactory resolution, some through admittedly short-term “band-aid” solutions, but most with a longer term fix available. Business confidence is not equally robust in each

¹⁷ If the California energy markets fail to clear, the holder of a transmission right usually gets a better position in the curtailment queue than a generator not holding a right.

ISO market (PJM appears to be the market with the fewest problems to date), but should increase as the markets mature.

Performance of Bilateral (Non-ISO) Markets. The largely bilateral markets, especially those in the Midwest, have experienced many potential reliability problems as evidenced by the frequency of curtailments under Transmission Loading Relief (TLR) procedures. These may also be attributable to the lack of independence of the system operator and market participants.¹⁸ Market participants have complained that they could not get access to the transmission system even when capacity appeared to have been available. As described below, Order 2000 requires the implementation of more efficient congestion management practices.

2.2 ORDER 2000 AND RTOs

A Regional Transmission Organization (RTO) is a transmission system operator that is independent of market participants, controls transmission facilities within a region of appropriate scope and configuration, and is responsible for operating those facilities to provide reliable, efficient and non-discriminatory service. All transmission owners must file a proposal to participate in a RTO or provide reasons for delaying or avoiding participation. Order 2000 explicitly notes that the designs for bid-based markets in the four ISOs operational before the year 2000 should form a basis for the design of RTO markets. However, the open architecture adopted in Order 2000 does not propose a single market model and offers sufficient leeway for further experimentation within the RTO design principles.

With respect to the unit commitment problem, the RTO has certain relevant functions. The RTO must have exclusive authority for maintaining short-term reliability. To fulfill this function, Order 2000 makes clear that the RTO requires knowledge of the operational status of generators and load.¹⁹

¹⁸ The curtailment of transactions in the presence of prices 10 to 100 times the annual average, due to TLRs, and voltage reductions, concurrent with power outages, indicate markets are not working in harmony with reliability constraints. For example, in the summer of 1999 the ECAR region with bilateral trading called 87 TLRs and the adjacent PJM ISO called three. For a general review of these complaints, see [6].

¹⁹ Such knowledge includes technical information supplied by generators such as ramp rates, upper and lower operating limits, whether the unit is running or not, start-up times and time between start-ups. In real-time and for day-ahead planning, the RTO must have information on generator injections and load withdrawals of energy in order to balance the system.

This includes control over interchange schedules, the authority to require redispatch of generation connected to the grid, and approval over scheduled outages.

The RTO will determine the required amount of each ancillary service and the location where the service is to be provided. It will also act as provider of last resort of ancillary services. That is, market participants can self-supply or purchase ancillary services from third parties, but the RTO must have the capability to provide any residual. The RTO or a third party unaffiliated with market participants must provide real time energy balancing.

With regard to transmission, the RTO must ensure development of market mechanisms for congestion management and must develop procedures to take into account parallel path flows. The RTO will sell physically feasible, short- and long-term, tradable transmission rights. The RTO may choose to expand the transmission system and/or invest in advanced technology to increase capacity.²⁰

Finally, the RTO is required to monitor for market power abuses and market design flaws. It should also evaluate and implement potential efficiency improvements in the markets it operates.

Beyond these requirements and guidelines, specific market designs are left to the RTO market developers (subject to the proviso that they not limit the RTO's ability to improve efficiency further). The remainder of this section discusses some issues about the conceptualization of the role of the RTO with respect to financial and physical transactions as well as the relationship of RTO-operated auction markets in relation to other energy markets. In addition, some pressing market design issues are reviewed, including pricing of reserves and inter-regional coordination. Section 3 then draws on the ISO experience and other sources to outline some principles for the design of day-ahead RTO markets.

Relationship Between Physical and Financial Transactions. An issue that has remained contentious in the preliminary design and operation of ISO markets is the relationship between physical and financial markets—specifically, the concern that the centralized ISO markets and nodal congestion pricing would inhibit development of the decentralized financial

²⁰ In the comments on Order 2000 [6], various policy suggestions were made regarding increasing transmission capacity, including overbuilding the transmission system (see Joskow comments) and/or investing in the high tech Flexible AC Transmission System (FACTS) and Wide-Area Measurement System (WAMS) to allow more robust competition to develop. The latter appears more promising because it promises not only more capacity and less greenfield construction, but also better system control (see EPRI and EEI comments).

markets.²¹ An important principle underlying the future RTO markets is that well-functioning physical markets promote robust financial markets. For our purposes, physical trades are trades that the RTO has registered as feasible, considering all other physical trades and required ancillary services. This includes bids into the ISO markets, bilateral transactions and self-schedules that have been cleared in the ISO day-ahead schedule (even though these day-ahead transactions are actually financial contracts until physical delivery). Financial trades are trades that are not physical trades, but take the form of forward contracts, futures contracts, or options contracts. They are not considered physical trades until they are confirmed as physically feasible by the RTO. Indeed, the RTO should be concerned primarily with physical market transactions; it would not operate purely financial markets and need not be involved in any financial markets unless the transaction goes to delivery.

Financial markets can and must exist in harmony and equilibrium with physical reliability markets. If not, the financial markets' ability to reduce risk is diminished. Multiple PXs and bilateral trading can fit easily into this market framework. Each PX would act as a single scheduler, submitting schedules and technical information for generation and load to the RTO.

The market design of the physical market should allow full, but optional interaction with financial markets. To become physical transactions, a market participant need only self-schedule, that is, submit quantities to the day-ahead market. If the transaction includes the necessary transmission rights and ancillary services, no charges will be assessed. This allows for fully hedged financial transactions. In addition, payment would be received for any additional service provided. Otherwise, the market participant will be billed for congestion, losses, and ancillary services caused by the bilateral transaction.

If not self-supplied, a bilateral trader can place price limits on what it is willing to pay for transmission and ancillary services. If the price limits are not met, the transaction will not be scheduled. If all voluntary adjustments are tried and reliability constraints are still not met, the transaction will be canceled in the day-ahead market. This gives ample time for parties to make adjustments. This cancellation avoids a potential TLR and the resulting schedule is very likely to be physically feasible.

²¹ One argument has been that uncertainty over nodal congestion prices, calculated hourly in real-time, increases the risk of bilateral deals concluded prior to the hour. Another is that some rules regarding three-part bids, in which the start-up payments made by the ISO are averaged over all electricity load in the system (e.g., in New York), effectively amounts to a subsidy to generators in the ISO auction market.

Physical markets provide real-time price signals and additional liquidity. Without good price signals from the physical markets, the financial markets can become unstable and encourage more speculation and less hedging.

Even though bilateral trading may be highly discriminatory (that is, sellers may charge different prices for the same delivered product to different buyers), the opportunity for buyers to participate in an efficient, nondiscriminatory RTO auction market will tend to discipline the bilateral market. RTO auction markets create options for all buyers and sellers and thereby allow for a light-handed approach to the regulation of these transactions. In sum, the benefits of well-designed RTO markets include lighter-handed regulation of financial markets, more liquidity, less gaming and risk, more visible prices, lower transactions costs, fewer curtailments, and compatibility with financial markets.

Reserves Markets. The establishment of efficient ancillary service markets is an ongoing market design challenge. As the cost of reliability increases and in the absence of a way to represent willingness to pay for ancillary services, the RTO system operator can relax reserve margins and transmission constraints. This is, in fact, written into the market rules in some ISOs; it remains a contentious issue, largely because it has involved system operator discretion that results in changes in market prices. The relaxation of these constraints increases the probability of a system failure; as such, it should be part of the operational parameters of the auction decided in advance of the day-ahead auction so that actions of the system operator are not seen as arbitrary.

To be effective, reserves must be able to respond to loads that need them. If transmission is congested between generation and load, reserves on the generation side are of little help to the load side. Both transmission and generation can be used to meet reserve requirements. They have substitute characteristics (strengthening transmission connections within a region can lower the total generation reserves needed in a region) and complementary characteristics (if the reserves for a load are not located at the same node, transmission capacity between the reserves and load will also need to be set aside). This set aside is similar to the capacity benefit margin (CBM) concept. The auction algorithm would set aside transmission capacity to allow reserves to respond. The modeling of this process is very similar to the modeling of the energy market itself. Prices for reserves would have a locational component and the transmission price would reflect the set aside transmission capacity.

Using this locational method of allocating reserves, it could appear that transmission capacity is being withheld. One method to deal with this is to have a “use or lose” requirement of transmission rights. The set aside “use”

of transmission for reserves markets would be considered a “use,” not withholding, and is under the control of the auction process and the operator. Dealing with the combination of congestion constraints and soft reserve constraints simultaneously requires operator independence and transparency in the market environment, as a means to promote trust in the market.

Over time, as demand becomes more responsive to price, generation reserve margins can decline as offers to reduce demand can substitute for reserves. Currently the costs of reserves are averaged among all end users. Some of these costs can be more directly assigned to specific generating units and individual customers. For example, a unit with a good reliability record should be responsible for a lower reserve margin or be charged less for reserves based on size and the historic probability of unit failure. Payments or discounts that differentiate more reliable from less reliable generators should be handled in the pre-day-ahead markets.

RTO and expanded inter-regional coordination. Spatial boundary conditions—also called the “seam” or interface problem—are becoming an important design issue as trading between ISOs increases. ISO coordination efforts on this matter are in the nascent stages. In terms of system representation, some ISOs have included a reasonably detailed representation of the interconnection with control areas outside the ISO as part of the boundary conditions. Approaches to the seam problem have included proposed interface auctions for inter-control area exchanges [16,17]. In day-ahead markets, there is some time to coordinate these interfaces offline via iterative trading rules. Research into this problem generally is in its infancy. A broad set of inter-regional market design and procedural issues is being examined by PJM, New York, New England, and Ontario, which have signed a memorandum of understanding on inter-regional coordination (on the need for such coordination generally, see [6]). Order 2000 anticipates that RTOs will assist in creating larger regional markets in which seams issues are resolved.

3. DESIGN PRINCIPLES FOR DAY-AHEAD RTO AUCTIONS

In the early phases of electricity market design and implementation, the various disciplines—notably economics and electrical engineering—have not undertaken adequate inter-disciplinary research or sufficient professional exchanges. For example, ISOs have several times misunderstood the incentive issues in electricity and ancillary service auction designs, as evidenced in the remedial actions and market re-designs described in Section

2, above. At the same time, market design has sometimes proceeded with the economics basically correct but without an adequate consideration of reliability and technical constraints. As a result, there is sometimes little understanding of what basic principles ought to underlie these complex markets. This section attempts to elaborate such principles for day-ahead markets. These principles can be compared with the market design principles in [14].

This section addresses the prospective RTO day-ahead market, which is defined as the market in which the initial bidding to provide energy and ancillary services for reliability, congestion management, and energy balancing takes place. As is done currently in ISO markets, this market would be conducted on the day prior to the dispatch day. The dual objectives of the day-ahead market are to achieve economic efficiency and ensure system reliability. The day-ahead market is a physical market where all expected balancing/ancillary services are scheduled. The design for the day-ahead market is discussed below and it assumes that there is a real-time market, in which adjustments are made to energy and ancillary services reflecting the differences between day-ahead expectations and real-time conditions.²²

The following is a list of recommended design principles derived from the foregoing analysis and experience to date, with short explanations and clarifications.

Principle 1: Maximize economic efficiency. The RTO auction objective is to maximize economic efficiency through voluntary market bids, bilateral transactions and self-scheduling, given the physical and reliability constraints.

Not all current ISOs adopt this principle. For example, the California ISO cannot adjust power schedules submitted to it by the California PX to improve economic efficiency. This principle requires that if market participants use the RTO markets, the resulting prices are efficient. The market-clearing procedure will balance the system and the purchase of ancillary services

²² The real-time market will not be examined in depth here. However, efficient market design requires that most principles be adhered to with respect to the relationship of the real-time and day-ahead markets. Bids should be submitted separately into the real-time market, and market prices based just on those bids. Deviations from the day-ahead market should pay the real-time price unless there is a reliability problem. If a bidder does not deviate from the day-ahead schedule, there are no additional costs to pay based on the real-time market. Finally, the market operator needs to keep the system in balance at the nodal level using bids to the extent possible.

using the bids it has received. If the financial markets are efficient, there may be few additional trade gains in this market and this market becomes a reliability check. Efficiency requires that prices be consistent (e.g., no price inversions due to market design flaws) and that arbitrage opportunities reflected in the bids be exhausted.

Principle 2: Ensure physical feasibility of market transactions and system reliability.

Without physical feasibility, reliability problems cannot be fully addressed. For ancillary services, the market design should require that generators committed to provide these services are located so that their capacity is available when and where they are needed.

Principle 3: Remove disincentives to market participation. Participation in an RTO market should involve low transaction costs and create minimal additional risks.

Minimal participation in the market is the submission of generation and consumption quantities (that is, bids which are taken at any price, or “at market”). Any unit dispatched should be guaranteed bid-cost recovery.

Principle 4: Bidding protocols should promote flexibility of participation. All market participants should be allowed, but not required, to submit multi-part bids that reflect short-term marginal costs. Market participants should be allowed to self-schedule, that is, allowed to submit quantity only bids.

This principle requires that all resources have the option to bid a reasonable approximation of their short-term marginal cost function, including start-up, no load, and energy costs (in addition to technical parameters, such as minimum and maximum load limits, ramp rates, and minimum shutdown time). Although a bid function will seldom serve as a perfect match for actual marginal (incremental or going forward) costs, a good approximation should be available.

This principle will require changes in some current unit commitment auctions, which allow only one-part energy bids (of course, in a multi-part bidding rule, generators can still submit one-part bids, by bidding zero for start up and no load). As discussed above, there are technical, financial, and economic reasons for adopting multi-part bidding.

While the multi-part bid allows for more accurate representation of marginal costs and thus, in the absence of market power, should result in a

more efficient solution, it also results in a non-convex supply function. In turn, this makes the market equilibrium and prices harder to derive. This complication can be addressed and made manageable with some simplifying assumptions about which generators are allowed to bid nonconvex costs, whether some parameters should be fixed for a specified period (such as ramp rates, maximum and minimum output), and what should be fixed in the bid function.

Demand bid functions are essentially the mirror images of generator bid functions but will not be discussed in detail here. Consumers need more explicitly defined contracts to participate properly in the market and should be allowed bid functions similar to the generators.

Principle 5: Make clear the distinction between financial and physical commitments. If accepted, bids are financially binding. If needed for reliability, bids are physically binding.

The RTO schedules are physical and financial commitments subject to liquidated damages. That is, if market participants deviate from their commitments, they are liable for making the affected market participants whole. Under emergency conditions, the RTO may exact other penalties for non-performance and/or issue perform-to-contract orders.

Principle 6: Minimize opportunities for arbitrage between different product markets.

This principle addresses generally the problems that arise when generators can bid into different product markets (energy and ancillary services) that are sequentially cleared; as discussed above, simultaneous markets eliminate opportunities for arbitrage.

If the market design allows opportunities for lower quality products to be priced higher than high quality products, then there may be cases where generating units are paid more for not generating than for generating. Preferably, this should not be the case. The New England ISO attempted to establish this principle administratively by proposing that the energy price always act as a cap for operating reserves prices. This proposal was rejected by the Commission; administrative measures should be at best transitions to market designs which can efficiently reach the same outcome. As discussed above, the simultaneous auction design will automatically allocate energy and reserves efficiently, but cannot guarantee the elimination of price inversions in the presence of market power.

Principle 7. Prices should be unbundled where possible to minimize averaging (i.e., socializing) of costs.

Averaged prices (whether for congestion, reserves or other costs) do not send the correct price signals for the entry of new generation. Uplift charges, the mechanism used for passing through costs of energy and ancillary services not covered by market prices, should be coupled with incentives for the RTO to minimize the use of such charges.

Principle 8. Market-clearing information should be made available as soon as possible. Fuller information on bids should follow with a suitable delay.²³

Market-clearing prices and quantities are the basic results of the bid acceptance process. They enable market participants and potential future market participants to assess the market and plan their businesses efficiently. They also allow market participants to spot and correct obviously erroneous bid acceptance and rejection decisions.

Disclosure of individual bids should be made eventually, but not immediately. Such disclosure will allow detection of subtle bid acceptance errors and it will also allow study of the market by independent analysts and market participants. It may lead to the exposure of the exercise of market power. Immediate disclosure of individual bids is undesirable because it might facilitate collusion by the market participants. Immediate disclosure might reveal information about market participants who wish to keep their costs confidential. After 6 months or a year, the information on individual bids has essentially no value for collusion and discloses little new about any bidder's current costs, but the information would have high value for auditing and independent analysis.

The auction software should be available to the market participant or public at a reasonable cost. Improvements to the software are desirable, and the best way to accomplish this is by making the software available with a set of test problems.

²³ The California ISO, PJM and New England appear to set the benchmark with updated prices every 5 minutes. At a minimum, prices and aggregate quantities should be available before the next round of bids with enough lead time to allow a reasonable response to the new information.

Principle 9: Minimize the incentives for market participants to engage in strategic behavior. The design should not favor market participants with market power.

Market designs can only imperfectly address structural sources of market power. An auction does not eliminate the ability of a large firm to withhold capacity profitably. Auctions have been devised to “pay-off” market power, but these require significant computation and may not be revenue sufficient.²⁴ Hence, absent a market design solution, the basic problem of structural market power has to be addressed using both structural remedies (vertical and horizontal dis-integration, encouragement of entry) and regulatory remedies, such as market power monitoring and mitigation. Of course, structural remedies may be wrapped up in market design issues. For example, an ISO may seek to promote rules on transmission interconnection for new generation which appear to favor incumbent generators.

While market design cannot necessarily mitigate structural market power, it can certainly exacerbate it; market design can also create opportunities for strategic behavior by generators other than the obvious large players. An example, discussed above, is the sequential clearing of energy and ancillary service markets without substitution in both the initial California and New England markets. Even a small generator can try to take advantage of shortages of certain types of reserves in this type of market to raise prices.

4. CONCLUSIONS

The considerable developments in the design of electricity markets over the past few years have provided the groundwork for the next generation of short-term markets. This chapter has emphasized that unit commitment models are now embedded in a variety of market contexts governed by an evolving regulatory framework that presents new requirements for modeling, including incorporation and understanding of market design issues. The open architecture promoted by the Commission allows for continued experimentation with RTO market design, but within parameters reflecting lessons learned heretofore from the ISO markets as well as the non-ISO bilateral markets. The market design principles presented in the paper are intended to reflect those lessons.

²⁴ The Vickrey-Clark-Groves (VCG) mechanism, which has been applied to electricity auctions by [15], is a method to elicit truthful bids from players with market power.

The other primary argument in the chapter is that a well-designed RTO day-ahead auction market with unit commitment complements “decentralized” financial markets. The computer does the work of resolving reliability constraints and will ensure that no offered trade gains are missed. Financial markets are free to create whatever innovative deals they can. The only restriction is that if they go to delivery they must satisfy any reliability constraints. Also, the opportunity for buyers to participate in an RTO auction market will tend to discipline the financial market. This will allow for lighter-handed regulation of financial transactions. The paper identified several other benefits of well-designed RTO markets: fewer curtailments, more visible prices, lower transactions costs, and less gaming and risk.

The open architecture also allows for continued progress in efficient pricing, so that causality has financial consequences—prime candidates are transmission rights and reliability. Over time, as price signals are sent and acted on in real-time, accurate pricing can allow the public good aspects of these markets to shrink in importance and the private good aspects to grow.

ACKNOWLEDGEMENTS

This paper reflects ongoing discussions among the authors and Carolyn Berry, Judith Cardell, Benjamin Hobbs, Thanh Luong, David Mead, William Meroney, and Roland Wentworth; it also benefited from the helpful comments of an anonymous reviewer. None of the afore-mentioned bear any responsibility for any errors or for any of the views expressed in the paper.

REFERENCES

1. P. Joskow and R.M. Schmalensee. *Markets for Power*. Boston: MIT Press, 1983.
2. F.C. Schweppe, M.C. Caramanis, R.E. Tabors, and R.E. Bohn. *Spot Pricing of Electricity*. Norwell, MA: Kluwer Academic Press, 1988.
3. H. Chao and H.G. Huntington, eds. *Designing Competitive Electricity Markets*. Boston: Kluwer Academic Press, 1998.
4. Federal Energy Regulatory Commission. Promoting Wholesale Competition Through Open Access Non-discriminatory Transmission Services by Public Utilities and Recovery of Stranded Costs by Public Utilities and Transmitting Utilities. Order No. 888, 61 FR 21,540, May 10, 1996.

5. Federal Energy Regulatory Commission. Open Access Same-Time Information System (formerly Real-Time Information Networks) and Standards of Conduct . Order No. 889, 61 FR 21,737, May 10, 1996.
6. Federal Energy Regulatory Commission. Regional Transmission Organizations. Order No. 2000, 89 FERC & 61,285, December 20, 1999.
7. C. Berry, B. Hobbs, W. Meroney, R. O'Neill, and W. Stewart. Understanding How Market Power Can Arise in Network Competition: A Game Theoretic Approach. Forthcoming in Utilities Policy.
8. S. Borenstein, J. Bushnell, and C.R. Knittel. Market Power in Electricity Markets: Beyond Concentration Measures. PWP-059, Univ. of Calif. Energy Institute, Berkeley, CA, Revised January 1999.
9. California ISO. Annual Report on Market Issues and Performance. Prepared by the Market Surveillance Unit, California Independent System Operator, June 6, 1999.
10. R.E. Bohn, A.K. Klevorick, and C.G. Stalon. Second Report on Market Issues in the California Power Exchange Energy Markets. Prepared for the Federal Energy Regulatory Commission by The Market Monitoring Committee of the California Power Exchange, March 9, 1999.
11. S. Borenstein and J. Bushnell and F. Wolak. Diagnosing Market Power in California's Deregulated Wholesale Electricity Market. PWP-064, Univ. of Cal. Energy Institute, Berkeley, CA, Revised, March 2000.
12. P. Cramton and J. Lien. Eliminating the Flaws in New England's Reserve Markets. Working Paper, Department of Economics, University of Maryland, College Park, MD, March 2, 2000.
13. J.E. Bowring, W.T. Flynn, R.E. Gramlich, M.P. McLaughlin, D.M. Picarelli, and S. Stoft. Monitoring the PJM Energy Market: Summer 1999. PJM Market Monitoring Unit, undated draft.
14. R.D. Wilson. "Design Principles." In [3].
15. B.F. Hobbs, M. Rothkopf, L. Hyde, and R.P. O'Neill. Evaluation of a Truthful Revelation Auction in the Context of Energy Markets with Nonconcave Benefits. J. Regulatory Econ., 2000, in press.
16. M.D. Cadwalader, S.M. Harvey, W.W. Hogan, S.L. Pope. Coordinating Congestion Relief Across Multiple Regions. PHB Hagler Bailly Inc., Navigant Consulting Inc., and J.F.K. School of Government, Harvard University, Cambridge, MA, October 7, 1999.
17. B.H. Kim and R. Baldick. Coarse Grained Distributed Optimal Power Flow. IEEE Transactions on Power Systems, 12 (2): 937, 1997.