REAL-TIME BALANCING OPERATIONS AND MARKETS: KEY TO COMPETITIVE WHOLESALE ELECTRICITY MARKETS

Eric Hirst
Consulting in Electric-Industry Restructuring
Oak Ridge, Tennessee 37830

April 2001

Prepared for
Edison Electric Institute
Washington, DC
Russell Tucker, Project Manager

and

Project for Sustainable FERC Energy Policy
Alexandria, VA
Terry Black, Project Manager

CONTENTS

Page

SUMMARY v

LIST OF ACRONYMS vii

1. INTRODUCTION 1
    FERC REQUIREMENTS 1
    MARKET FUNCTIONS AND GOALS 2
    PROJECT STRUCTURE 4

2. UNIT COMMITMENT AND DISPATCH 7
    UNIT COMMITMENT 8
    DISPATCH 9
    AREA CONTROL ERROR 11
    OPERATIONS WITH COMPETITIVE MARKETS 13

3. COMPETITIVE-MARKET PRICES 17
    DAY-AHEAD AND REAL-TIME PRICES 17
    REAL-TIME HOURLY AND INTRAHOUR PRICES 19
    SUMMARY 20

4. EXAMPLES 23
    INTRODUCTION 23
    RAMPRATE LIMIT 26
    LOW-OPERATING LIMIT 28
    STARTUP COSTS 29
    ENERGY-LIMITED UNITS 31
    RAMPRATE LIMITS AND MULTIPLE INTERVALS 33
    IMPORTS AND EXPORTS 34
    INTERVAL VS HOURLY SETTLEMENTS 36
    INTERMITTENT RESOURCES 38

5. CONCLUSIONS 43

ACKNOWLEDGMENTS 44

REFERENCES 45

APPENDIX: ISO EXPERIENCES AND RTO PLANS 49
NEW YORK 49
CALIFORNIA 51
ISO EXPERIENCES 52
EXAMPLES OF ISO PROBLEMS 54
RTO PLANS 56

SUMMARY

Electricity production and consumption must occur at essentially the same time. Therefore, real-time (minute-to-minute) operations and the associated markets and prices are essential ingredients of a competitive wholesale electricity industry. In addition, these intrahour markets are the foundation of all forward markets and contracts, including hour- and day-ahead markets, monthly futures, and bilateral contracts. Finally, these intrahour operations maintain system reliability by ensuring that enough and the right kinds of supply and demand resources are available when needed.

Because of various load, generation, and transmission factors, balancing generation to load on a minute-to-minute basis is complicated. Loads are volatile, both from hour to hour and from minute to minute during the morning rampup and evening dropoff. Generators differ substantially in their costs of electricity production. In addition, generators have various idiosyncratic characteristics, such as maximum and minimum output levels and maximum ramprates, that limit their ability to respond rapidly to changes in system load or generation. Finally, transmission characteristics affect the real-time balancing function because of congestion and sudden transmission outages. These factors can lead to dramatic and rapid changes in electricity prices, including occasional negative prices when generators pay someone to take their output (Fig. S-1).

The early years of operations by independent system operators (ISOs), based on the experiences in New England, New York, and California, show how difficult it is to translate the theory and initial design of competitive markets into ones that work efficiently. These ISOs have been plagued with various startup problems that artificially raise electricity costs to consumers, implicitly encourage strategic bidding by some generators, do not sufficiently discipline generator market power, sometimes yield insufficient resources, and impair reliability. Fortunately, the ISOs have been quick to identify and remedy flaws in their initial market designs. On the other hand, the ISOs have done a poor job of documenting these problems and their resolutions.

This report is primarily a primer on how such intrahour operations and markets should work. It demonstrates these principles through several examples. These examples deal with generator ramprate limits, low-operating limits, startup costs, and other generator characteristics. Other examples show how energy-limited (e.g., hydro) units differ in their bidding and operations from capacity-constrained (e.g., thermal) units, how the consideration of multiple time intervals affects operations and pricing, how generators located outside the control area are treated differently from those within the control-area boundary, how interval

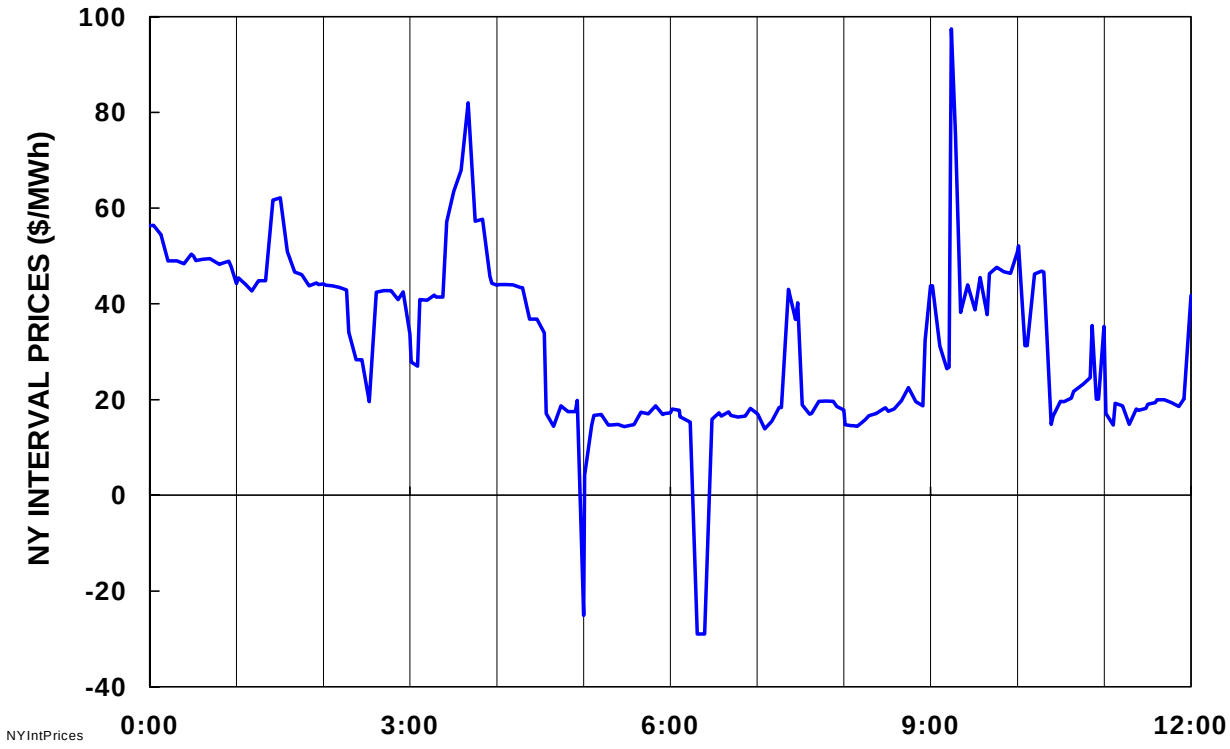


Fig. S-1. Intrahour prices for the New York ISO West zone for a 12-hour period in September 2000.

pricing combined with hourly settlements can encourage generators to ignore dispatch signals, and how intermittent resources (such as wind) affect control-area operations.

Although U.S. wholesale competitive markets today suffer from a variety of problems, there is reason to be optimistic. Ultimately, the ISOs (and, later on, RTOs) will identify and fix the problems within their market structures, and they will adequately document their problems and the associated resolutions so that their market participants and the designers of other systems can learn from past mistakes. Ultimately, efficient real-time markets should allow reliability councils and system operators to largely replace command-and-control rules with market signals (i.e., prices that vary rapidly in response to changes in system security). These changes should lower the costs of maintaining reliability; deploy supply and demand resources more efficiently; and guide investments in new generation, transmission, and demand-side resources.

LIST OF ACRONYMS

ACE	Area control error
BME	Balancing-market evaluation
CPS	Control Performance Standard
CT	Combustion turbine
FERC	U.S. Federal Energy Regulatory Commission
ISO	Independent system operator
LBMP	Locational-based marginal price
LOL	Low-operating limit
MCP	Market-clearing price
NERC	North American Electric Reliability Council
NYSEG	New York State Electric & Gas Company
NYISO	New York Independent System Operator
PX	Power exchange
RTO	Regional transmission organization
SC	Scheduling coordinator
SCD	Security-constrained dispatch
SCUC	Security-constrained unit commitment

INTRODUCTION

Electricity is the ultimate real-time product, with its production and consumption occurring at virtually the same time. This simultaneity of production and consumption is a consequence of the fact that electricity cannot be easily stored. Because of the physics of bulk-power electric systems, system operators must dispatch generation up or down to follow minute-to-minute changes in load and generation output. Absent such near-real-time balancing, electrical systems would be highly unreliable with frequent and severe outages.

Because electricity production must be increased or decreased frequently, hourly energy markets are insufficient to maintain reliability. Because of transmission-network externalities and the speed with which decisions must be made and implemented, it is not possible to let the market determine the rules and products. "The market process itself is a natural monopoly that requires some degree of social design and regulation to assure that all traders have equal and non-discriminatory access to the market and that the interests of affected parties not directly acting in the market—particularly small consumers—are protected" (Ruff 2000). Therefore, system operators manage intrahour imbalance-energy markets with intervals of typically 5 or 10 minutes. (In addition, system operators acquire and deploy ancillary services to maintain reliability in real time.) This report focuses on these intrahour operations and markets, the actions that system operators take to maintain the necessary generation-load balance and the markets they use to acquire the incremental and decremental energy needed to maintain that balance.

Although most energy is scheduled in forward markets (from years and months ahead through an hour ahead), some imbalances unavoidably occur in real time because of various generation, load, and transmission factors. Generation factors include forced outages, ramprate limits, generators not following their schedules accurately, the use of some generation to provide ancillary services, and the intermittent nature of some generation (e.g., wind). Load factors include sudden changes in weather conditions that affect space-conditioning electrical loads, weather-forecasting errors, and intrahour load changes. Transmission factors include local and regional congestion, unscheduled (parallel) flows, and forced outages. All these factors, singly and in combination, require the system operator to have access to generation output that it can move up or down from interval to interval.

FERC REQUIREMENTS

The Federal Energy Regulatory Commission (FERC 1999), in its landmark Order 2000 on regional transmission organizations (RTOs), recognized the importance of these markets. FERC wrote:

> … an RTO must ensure that its transmission customers have access to a real-time balancing market that is developed and operated by either the RTO itself or another entity that is not affiliated with any market participant. We have determined that real-time balancing markets are necessary to ensure non-discriminatory access to the grid and to support emerging competitive energy markets. Furthermore, we believe that such markets will become extremely important as states move to broad-based retail access, and as generation markets move toward non-traditional resources, such as wind and solar energy, that may operate only intermittently.
>
> … real-time balancing markets are essential for development of competitive power markets. Therefore, although we will give RTOs considerable discretion in how they operate real-time balancing markets, we will not allow implementation of such markets to be discretionary.

In essence, FERC ordered RTOs to create and operate real-time markets to maintain short-term reliability by relying primarily on markets to acquire reliability services and to foster competition throughout the electricity industry.

MARKET FUNCTIONS AND GOALS

Largely because of the simultaneity requirement noted above, spot (real-time) electricity prices are highly volatile. Other factors affect price volatility:

- Generators differ substantially in their costs to produce electricity [e.g., the running costs for hydro and nuclear units are typically well below $10/MWh, while the cost for a combustion turbine (CT) might be $100/MWh or more].

- System loads vary substantially intrahour during the morning rampup and evening rampdown.

- Sudden generator outages, transmission outages, extreme weather conditions, and other events can trigger unexpected imbalances between generation and demand; rebalancing the electrical system can be expensive.

- Intertemporal constraints (e.g., ramp and acceleration rates) limit generator flexibility so that the least-expensive units cannot always adjust to meet rapidly changing loads.

In addition to the reliability requirements, real-time operations are an essential ingredient of competitive wholesale electricity markets. Although most electricity is bought and sold long before consumption (including bilateral contracts and monthly block-forward markets, day-ahead markets, and hour-ahead markets), real-time prices serve as a benchmark against which all forward markets settle. The prices agreed to in forward markets by buyers and

sellers are based on their expectations for the real-time prices that, ultimately, will occur. If market participants expect real-time prices to be competitive and efficient, forward prices will also be competitive and efficient. On the other hand, if market participants expect real-time prices to be biased by market-power abuses, inefficient market rules, or poor operating practices by the system operator, then forward prices will not reflect the societal value of the electricity being traded. The reverse is also true: if forward markets are inefficient, costs may be higher in real-time operations, and reliability may be more difficult to maintain. Morey (2001) discusses the design of these forward auctions.

Finally, real-time prices motivate long-term investment decisions. These decisions affect changes to existing generating units to extend their lifetimes or repower them. They affect investments in new generating units. And they affect investments in metering, communications, and control technologies that permit retail customers to modify the timing of their electricity use in response to volatile electricity prices (Hirst and Kirby 2001).

Although electric-industry market participants debate the relative merits of decentralized bilateral arrangements vs centralized markets, all agree that, in real time, centralized control is required. As Ruff (2000) notes, decentralized energy markets cannot fully anticipate the real-time grid complexities that might occur. Therefore, the system operator "must have some way to determine, motivate, and compensate, in real time, a reasonably efficient set of physical actions that may differ significantly from those implied by decentralized market trading."

The design of bulk-power markets includes the number and types of different products (energy, capacity, and ancillary services); bidding and scheduling processes; the number of and relationship among forward and real-time markets; market-clearing and settlement rules; and types of congestion management and transmission rights. This project focuses on a subset of these issues—the design and operation of intrahour balancing markets—because of their importance and complexity and because of the problems ISOs have had in making them efficient and attractive to market participants. Although the markets for energy, ancillary services, and transmission (congestion) are closely coupled and should be considered in an integrated fashion, I here treat energy only. I ignore congestion and ancillary services because energy-only markets are sufficiently complicated. For similar reasons, I do not discuss retail-load participation in such markets.

The primary goal of intrahour operations and markets should be to maintain reliability at least cost. Attaining this goal requires the system operator to balance generation against load to minimize the costs of energy, congestion, losses, and contingency management. (Whether the system operator should minimize the cost to consumers or the suppliers' bid costs is a contested issue.) A well designed market encourages suppliers to bid their costs accurately because bidding otherwise would lower their profits. With suppliers acting as price takers (absent market power), market designers can avoid the use of artificial penalties that increase overall costs by requiring individual market participants to overcomply with dispatch orders. Such a system would ensure that suppliers receive revenues that at least equal their offered

prices and that these prices are stable and predictable (e.g., when demand exceeds supply, prices rise and vice versa).

In addition, intrahour markets should be equitable (e.g., treat market and bilateral transactions fairly) and should avoid the use of penalties that are not cost based. These markets should also be transparent (i.e., easily understood by market participants) so the participants know how their bids are used and how prices are set.

PROJECT STRUCTURE

To conduct this project, I contacted staff of the operational ISOs and of several market participants (especially the generators). In addition, I reviewed the literature for relevant articles and searched the ISO websites for market rules and operating procedures. The results obtained from conversations with industry experts and the literature search were disappointing. Obtaining detailed information in today's dynamic environment is difficult because ISO staff are very busy. The market participants with whom I talked had many concerns about current ISO operations and rules but could provide few specifics. The literature contains several relevant papers, but these generally deal more with theory than with practice and provide few details on the practice and problems of intrahour operations and markets.

Although the ISO websites contain a wealth of information, many of the procedures that purport to describe these markets and operations are quite general. For example, the New York ISO (1999a and b) published a *NYISO Transmission and Dispatching Operations Manual* and a *NYISO Day Ahead Scheduling Manual*, neither of which contains sufficient detail to truly understand how the ISO operates its intrahour energy market. The comparable PJM publications (2000a, b, and c)—*PJM Manual for Pre-Scheduling Operations*, *PJM Manual for Scheduling Operations*, and *PJM Manual for Dispatching Operations*—contain even less detail than the New York manuals.

I did find three documents that were enormously helpful, including a New York Department of Public Service (2000) review of the New York ISO, an operational audit conducted by PricewaterhouseCoopers (2000a) of ISO New England, and a PricewaterhouseCoopers (2000b) audit of the California ISO. These reports provided considerable detail and discussions of current problems in ISO operations.

Because of the limitations of the available materials, I was unable to address all the topics I considered relevant (Table 1). In particular, I could not obtain sufficient details on the workings of the individual ISOs to draw conclusions on the key features of successful operations. As the PricewaterhouseCoopers (2000a) audit of ISO New England notes, "Detailed procedures should be developed and maintained in adequate detail to support the day-ahead and real-time operations of ISO-NE and guide employees in their daily activities. ISO-NE should increase the transparency of its operations through more timely communication of information to market participants."

To a large extent, this report is a semitechnical primer on some of the key issues that must be addressed in the design and operation of intrahour markets. I present several examples to illustrate these key issues. The material presented here should help market participants understand how real-time operations work and how intrahour (and therefore hourly) prices are determined. The material should also help ISO staff understand some of the policies and specifics of how other ISOs manage their real-time operations and markets.

The remainder of this report is organized as follows. Chapter 2 explains the unit-commitment and economic-dispatch processes for vertically integrated utilities; it also briefly discusses these processes for RTOs, with additional detail deferred to the Appendix. Chapter 3 reviews data from the four U.S. operational ISOs on electricity prices in day-ahead, real-time hourly, and intrahour markets. Chapter 4 presents several examples to illustrate the complexities associated with efficient design and operation of intrahour energy markets. Chapter 5 summarizes the results of this study. The Appendix explains the unit-commitment and dispatch processes employed in New York and California, briefly reviews some of the problems the existing ISOs have experienced in making their real-time markets work well, and discusses the paucity of information on these markets in the recent RTO filings with FERC.

Table 1. Critical issues related to real-time operations and markets

- Should unit commitment (resource scheduling) be done by individual suppliers, the RTO, or both?
- How do installed-capacity requirements and markets affect real-time operations and prices? What obligations might designated capacity resources face in real time?
- How do reliability-must-run resources affect real-time operations and markets?
- How are resources dispatched, by whom, and how are they compensated?
- What is the optimal time interval (e.g., 5, 10, or 15 minutes) for dispatch and price setting?
- Should the RTO establish a single market-clearing price (MCP) in each interval or pay each winning resource its bid price?
- Should the real-time market pay for energy only or pay also for maneuverability (e.g., ramp and acceleration rates)?
- How do resource constraints (e.g., upper and lower operating limits and ramprates) determine which resources can set the MCP and which cannot and why?
- Should intrahour prices be determined ex ante or ex post? If prices are set ex ante, what is the basis for the value?
- Are exports and imports treated differently from internal transactions? What rules govern interchange scheduling (e.g., the number of schedule changes per hour permitted and maximum ramprates for schedule changes)?
- To what extent does the RTO make short-term forecasts of load and generation, how far into the future (e.g., 10 minutes to 24 hours), and how does the RTO use these forecasts? Should the RTO commit and dispatch resources on the basis of expected future conditions (i.e., beyond the next interval)? Who pays for errors associated with these RTO decisions?
- Should the RTO publish prices and let demand and supply respond to the price signal, or should the RTO dispatch resources up and down based on supplier bids?
- If the RTO explicitly dispatches resources, should uninstructed deviations be treated differently, in terms of payment or penalties, from instructed deviations? What about a resource's failure to follow instructions?
- Under what reliability circumstances should the RTO go "out-of-market" for resources? What should set the payment to these resources?
- Under what conditions, if any, are penalties appropriate, and for what kinds of behavior? What determines the magnitude of the penalty? Should penalties apply to generation only or to loads also?
- How, if at all, should capacity assigned to ancillary services (especially the reserve services) be incorporated into real-time operations and markets? Should the capacity assigned to contingency reserves be set aside and used only when a major outage occurs? Or should such reserves be used routinely whenever it is economic to do so, as long as sufficient capacity is available to meet contingency-reserve requirements?
- How should intermittent resources (e.g., wind) be treated (compensated for energy deliveries) in real-time operations and markets? Should they be treated differently from volatile loads?
- Can retail loads participate in real-time markets? If so, how?

UNIT COMMITMENT AND DISPATCH

This chapter begins by explaining the unit-commitment and dispatch processes used by vertically integrated utilities as background for how these issues can be addressed in the new restructured environment. I begin with traditional utility operations because they address the same technical and economic issues (without the complexities of competitive markets) as those fundamental to restructured electricity markets.

Although this project focuses on intrahour operations, the discussion must begin earlier in time. Indeed, decisions made years ahead about the construction of new generating units and long-term fuel-supply contracts affect the options that are available in real time and their costs. For purposes of this discussion, however, I begin with the day-ahead unit-commitment (scheduling) process. Unit commitment is the process a utility goes through in deciding which units to operate the following day, and when to turn these units on and off. Such a process is necessary because electricity use varies throughout the day, often by a factor of two or more. This variation in electricity use argues for the operation of different units at different times during the day to minimize the overall costs of electricity production.

The second process discussed here is dispatch, the allocation of the generating units online or that can be turned on within a few minutes to meet current energy requirements. The dispatch process identifies the least-cost mix of generating units to meet demand, where cost is defined as the variable fuel plus operations and maintenance expenses. As Wood and Wollenberg (1996) point out, the dispatch problem is much simpler than the unit-commitment problem. Indeed, dispatch is a subset of unit commitment:

> The economic dispatch problem *assumes* that there are N units already connected to the system [online]. The purpose of the economic dispatch problem is to find the optimum operating policy for these N units. … On the other hand, the unit commitment problem is more complex. We may assume that we have N units available to us and that we have a forecast of the demand to be served. The question that is asked in the unit commitment problem area is approximately as follows. Given that there are a number of subsets of the complete set of N generating units that would satisfy the expected demand, which of these subsets should be used in order to provide the minimum operating cost?

UNIT COMMITMENT

Vertically integrated utilities typically run their unit-commitment optimization computer programs the afternoon of the day before operations. These large, complicated computer programs accept as inputs detailed information on the characteristics of the individual generating units that are available to provide electricity on the following day. These characteristics include current unit status, minimum and maximum output levels, ramprate limits, startup and shutdown costs and times, minimum runtimes, and unit fuel costs at various output levels. In addition, the operations planner inputs to the model the utility's day-ahead forecast of system loads, hour by hour, as well as any scheduled wholesale sales or purchases for the following day. Finally, the inputs include details on the characteristics of the transmission system expected for the operating day (in particular, any lines or transformers out of service for maintenance).

The optimization model is then run with all these inputs in an effort to identify the least-cost way to meet the following day's electricity demands while maintaining reliability. The reliability requirements include the ability to withstand the loss of any single generation or transmission element while maintaining normal service to all loads. The optimization model performs two functions in its search for a least-cost solution. First, it tests different combinations of generating units that are available and, therefore, could be scheduled to operate the following day (i.e., the times each unit will start, operate, and then be turned off). Second, given the units that are online and operating during any hour, it selects the least-cost mix to meet that hour's loads.

Solving this optimization problem is complicated because of all the intertemporal constraints that generators have. For example, one unit may be relatively cheap to operate (in terms of its variable costs, expressed in $/MWh) but may have relatively high startup and noload costs (expressed in $/startup and $/hour, respectively), while another unit has just the opposite characteristics. Which unit to commit depends on how many hours it is expected to operate the following day. In addition, the unit-commitment solution must respect system constraints, which include contingency-reserve requirements and the transmission constraints mentioned above. Finally, the optimization model must consider many different combinations of generating units that could meet the hour-by-hour loads during the day.

Because these problems are very difficult mathematically, the solutions are approximate. As a consequence, the final solution may not be exactly least cost. For a vertically integrated utility, this approximation is not a problem because its profitability depends on its entire portfolio of generating units, not on the performance (operation) of only one or two units. For the customers of such a utility, the differences among solutions are also generally inconsequential, because the differences in total costs between near-optimal solutions is small.

The unit-commitment program may be run several times during the operating day, especially if conditions change materially from the time the day-ahead run was made. Such

changes might include the loss of a major generating unit, the return to service of a large generator that was offline, the loss of a transmission line, or a change in system load caused by unexpected weather changes.

DISPATCH

Once generators are committed (turned on and synchronized to the grid), they are available to deliver power to meet customer loads and reliability requirements. Utilities typically run their least-cost dispatch model every five minutes or so. This model forecasts load for the next 5-minute interval and decides how much additional (or less) generation is needed during the next interval to meet system load.[*] The model may look ahead several intervals (up to an hour or more) to see if any quick-start units (e.g., CTs and hydroelectric units) should be turned on to meet projected demand over the next several intervals. The model then selects the least-cost combination of units that meet the need for more or less generation during the next intrahour interval. This combination must respect the constraints of each generator, including minimum and maximum operating levels and ramprates.

Utilities typically model the fuel costs[#] of their generators as a polynomial:

$$\text{Fuel cost (\$/hr)} = a + b \times \text{MW} + c \times \text{MW}^2 + \dots , \quad (1)$$

where a, b, c, … are constants and MW is the unit's output. The constant a represents the unit's noload cost, the hypothetical hourly cost to keep the unit on while producing no electricity. The b and c constants show how fuel costs increase with increasing unit output. The top of Fig. 1 shows a typical fuel-cost curve for a 500-MW unit with a lower-operating limit of 125 MW.

Dividing both sides of Eq. 1 by the unit's output yields the *average* fuel-cost curve shown in the bottom of Fig. 1. The average fuel cost (as well as average heat rate in Btu/kWh) is typically highest at the lowest operating point and lowest at 90 to 95% of the maximum operating point. The bottom of Fig. 1 also shows the *incremental* fuel-cost curve, defined as the first derivative of the equation above:

$$\text{Incremental fuel cost (\$/MWh)} = b + 2 \times c \times \text{MW} + \dots . \quad (2)$$

The typical incremental-cost curve increases with unit output and crosses the average-cost curve at the point of minimum average fuel cost (440 MW in the bottom part of Fig. 1).

---

[*]In addition, the dispatch model will seek to return the units providing the regulation ancillary service to their midpoints so these units are ready to provide the full range (up and down) of this service (Hirst and Kirby 2000).

[#]The dispatch process also considers variable operations and maintenance expenses. These costs can be included in Eq. 1 by appropriately increasing the b coefficient.

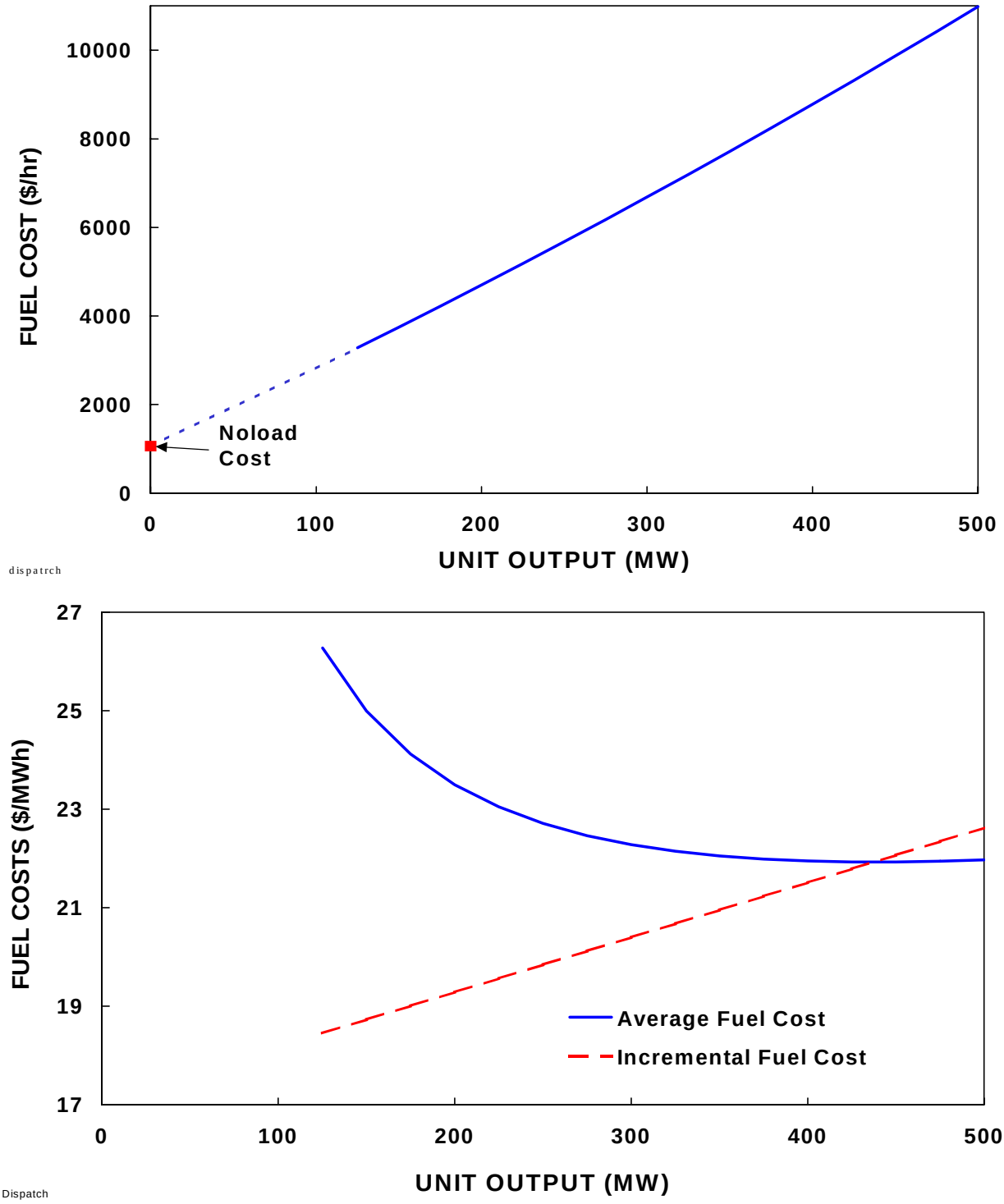


Fig. 1. **Variable fuel costs for a hypothetical generator. The top curve shows hourly fuel costs as a function of unit output. The *Y*-axis intercept (small square) is the noload cost, the theoretical cost to run the unit were it able to operate while producing no electricity. The bottom graph shows the average and incremental fuel costs per MWh as a function of unit output.**

In deciding whether to turn a unit on or not (the unit-commitment decision), the utility uses the average-cost curve as well as the unit's startup cost and minimum runtime. Once the decision is made to turn a unit on, however, decisions on its optimal output level are based solely on incremental costs.

Assume, as an example, that the 500-MW unit in Fig. 1 has a startup cost of $12,500 (equivalent to $25/MW). If the expected hourly prices for hours 12 through 17 are $24, $27, $32, $36, $26, and $23/MWh and are below $18 for the other 18 hours, it is worthwhile to turn the unit on and run it at full output for these six hours because these prices are all higher than the $22.6/MWh incremental cost at full output. The fuel cost is $65,900 and the revenues are $84,000. Subtracting the startup cost leaves operating income of $5,600.

If, however, the price in hours 14 and 15 was only $28 (instead of $32 and $36), the unit would not collect enough money from the sale of energy to cover both its fuel and startup costs. In that case, the day-ahead decision would be to leave the unit offline.

In real-time, if the unit was on and operating for hours 12 through 17, what should it do in hour 18 if the spot price drops to $21? Although this price is below the average-cost curve (bottom of Fig. 1), it is equal to the incremental cost at an output of 350 MW. If the unit's owner wants to keep the unit online (perhaps to capture earnings later that evening when prices might, once again, be higher), it should operate the unit at 350 MW. Otherwise, the unit should be shut down to save the noload cost.

AREA CONTROL ERROR

Fundamental to understanding the nature of generation-load balance is area control error (ACE). Control areas seek to minimize any adverse effect they might have on other control areas within its Interconnection by minimizing their ACE (North American Electric Reliability Council 1999). The ACE equation, in slightly simplified form, is:

$$ACE = (I_A - I_S) - 10\beta(F_A - F_S)\,, \tag{3}$$

where I refers to the algebraic sum of all power (MW) flows on the tielines between a control area and its surrounding control areas, F is the Interconnection frequency (Hz), A is actual, S is scheduled, and β is the control area's frequency bias (MW/0.1Hz). (Frequency bias is the amount of generation needed to respond to a 0.1 Hz change in Interconnection frequency. It is usually set equal to the supply-plus-load response of a control area to a change in Interconnection frequency.) The first term shows how well the control area performs in matching its schedules with other control areas (i.e., how well it matches its generation plus net incoming scheduled flows to its loads). The second term is the individual control area's contribution to the Interconnection to maintain frequency at its scheduled value (usually 60 Hz). Thus, ACE is the instantaneous difference between actual and scheduled interchange,

taking into account the effects of frequency. In essence, ACE measures how well a control area manages its generation to match time-varying loads and scheduled interchange.

To follow minute-to-minute variations in load, system operators use their automatic-generation-control systems to dispatch those generators providing regulation. The 5- or 10-minute economic dispatch discussed here is used to move generators up or down to follow trends in intrahour loads and to return the units on regulation to the midpoints of their range. However, generation and load need not (indeed cannot) balance instantaneously. NERC's Control Performance Standard (CPS) 1 and 2 determine the amount of imbalance that is permissible for reliability.

CPS1 measures the relationship between ACE and Interconnection frequency on a 1-minute average basis. For example, when frequency is above its reference value, undergeneration benefits the Interconnection and leads to a "good" CPS1 value. CPS1, although recorded every minute, is reported and evaluated on an annual basis. See NERC's (1999) Policy 1 - Generation and Control for additional detail on CPS1.

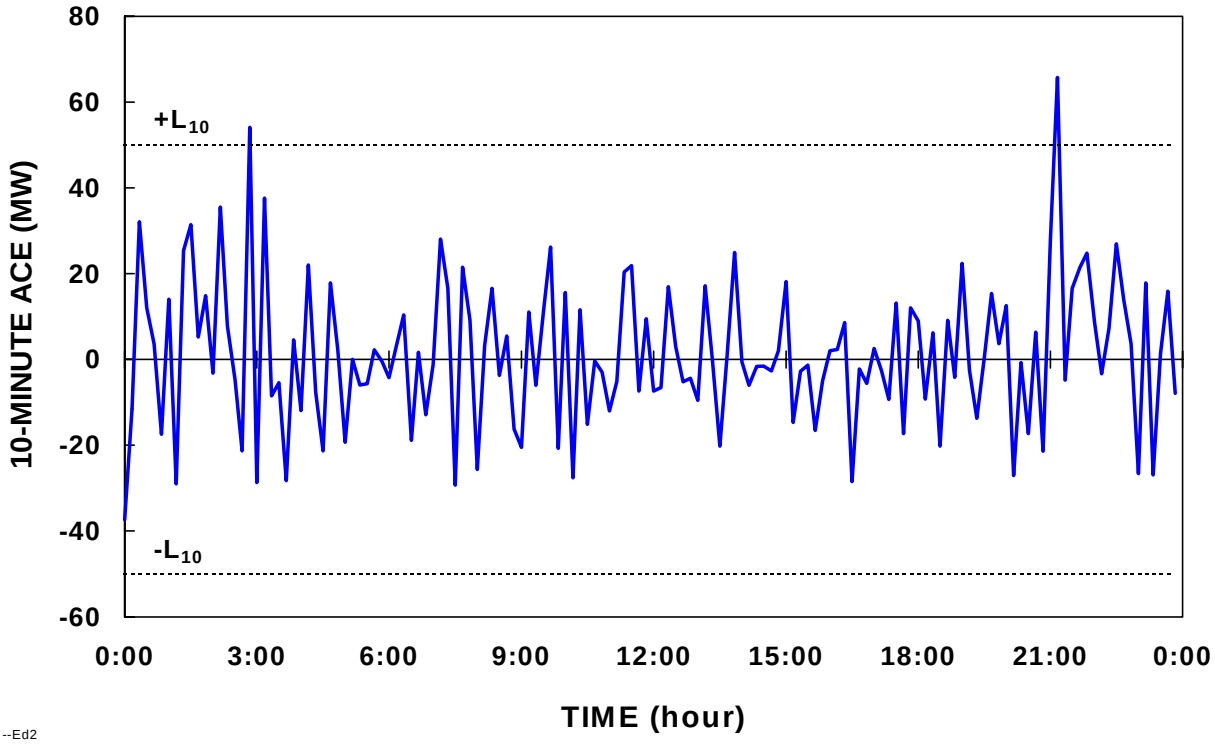
NERC's Policy 1 defines CPS2 as follows:

> The average ACE for each of the six ten-minute periods during the hour (i.e., for the ten-minute periods ending at 10, 20, 30, 40, 50, and 60 minutes past the hour) must be within specific limits, referred to as L10. See the Performance Standard Training Document, Section B.1.1.2 for the methods for calculating L10. … Each CONTROL AREA shall achieve … CPS2 compliance of 90% … .

The 90% requirement means that a control area can have no more than 14.4 CPS2 violations per day (10% of the 144 10-minute intervals), on average, during any month.

Figure 2 shows one control area's 10-minute average ACE values for a full day. This control area easily met the CPS2 criterion, experiencing only two violations. (CPS2 is a monthly, not a daily, measure of control-area performance.) The average ACE of 1.2 MW (equivalent to a total overgeneration of 29 MWh for the entire day) is tiny compared with this control area's average load of 11,000 MW. The average of the absolute values of ACE during each interval was 13 MW for this day.

This example shows that maintaining reliability does not require a control area to *exactly* balance generation to load each and every minute. Small imbalances are generally permissible, as are occasional large imbalances. Both CPS1 and 2 are statistical measures of imbalance, the first a yearly measure and the second a monthly measure.



--Ed2

Fig. 2. **Area-control error for each 10-minute interval of one day. Only two CPS2 violations occurred on this day, both involving overgeneration.**

OPERATIONS WITH COMPETITIVE MARKETS

As Wood and Wollenberg (1996) write, life is much more complicated when separate entities own and operate generation, transmission, and system control: "With only a single, integrated electric utility operating both the generation and transmission systems, the local utility could establish schedules that minimized its own operating costs, while observing all of the necessary physical, reliability, security, and economic constraints. With multiple parties in the bulk-power system (i.e., the generation and transmission system), new arrangements are required. The economic objectives of all the parties are not identical, and, in fact, may even be in direct (economic) opposition." Instead of optimizing resources within a single entity (the vertically integrated utility), the optimization takes place across the entire market.

The primary difference between RTO and vertically-integrated-utility unit commitment and dispatch is that the RTO owns no generation resources. As a consequence, it must sign bilateral contracts, operate markets, or both to acquire the generation it needs to maintain reliability and match generation to load in near-real time. Because of this split between generation and system control, risk allocation is a major policy issue in designing competitive electricity markets. Who (competitive suppliers or the RTO) should make unit-commitment decisions and bear the risks if these decisions turn out to be suboptimal in real time? Similarly,

should suppliers bid into energy markets with simple energy-only bids, or should they use multipart bids with separate prices for startup, noload, and energy costs?

The unit-commitment model illustrates well the complexities associated with the deintegrated operations of an RTO. As noted above, because the unit-commitment problem is very complicated, the solutions it produces may not be exactly least cost. Although this does not cause problems for vertically integrated utilities, it can do so for independent power producers. The political and market implications of choosing one unit over another can be dramatic in a competitive world when the choice can mean profit or loss for the chosen or skipped unit. This situation is quite unlike the historical one in which the entire portfolio of generation was owned by a single entity, the vertically integrated utility.

In addition, the accuracy of the unit-commitment solution depends strongly on the accuracy of the inputs to the model. Recall the old computer adage: garbage in, garbage out! While the traditional utility had no incentive to provide misleading inputs to its unit-commitment model, the same may not be true for competitive power producers. Indeed, each producer, seeking to maximize its profits, may modify its inputs to the RTO to "trick" the unit-commitment model into (inappropriately) selecting its unit(s) at high prices and profits. The Appendix section on ISO problems gives one such example from ISO New England.

Additional complications occur with the dispatch process. Energy bids can be one- or multi-part. With one-part bids, the suppliers must integrate startup and noload costs into the incremental energy bids. In the three-part structure (i.e., startup, no load, and energy costs), these characteristics are bid separately and are used in a centralized model to find the least-cost solution that guarantees recovery of all stated costs for units that are committed. In the former case, suppliers bear the risks of recovering these costs; in the latter case, the RTO may guarantee the recovery of bid costs for units that are scheduled to run.

The existing U.S. ISOs differ in their treatment of these issues, with considerable similarity among the processes used by the New York ISO, ISO New England, and the PJM Interconnection and substantial differences between these three ISOs and California.[*] The greatest difference between the Northeastern and California approaches is in unit commitment. In California, neither the ISO nor the Power Exchange (PX) conducts a centralized unit commitment. Individual suppliers are responsible for deciding how to bid their resources into the various energy and ancillary-services markets and are responsible for any unrecovered costs associated with those decisions. Unrecovered costs might include those associated with startup and shutdown as well as operation during certain hours when the MCP is below the variable cost of the unit. PJM, as a contrast, uses its day-ahead load forecast to schedule units for each of the 24 hours in an effort to minimize total operating costs across the control area. PJM

[*]This report does not deal with the problems California has been experiencing since late May 2000. In particular, the California Power Exchange stopped operation of its day-ahead and day-of markets at the end of January 2001 in response to a FERC (2000e) order.

guarantees that any generator it schedules and operates will not lose money during that day. In other words, if energy revenues do not exceed the unit's stated startup and noload costs as well as the variable operating costs, PJM will make up the difference through an uplift charge paid by all electricity consumers. [An uplift charge is a cost collected from customers on the basis of their hourly energy use ($/MWh). Such costs might include congestion management (redispatch of generators) as well as startup and noload.]

A second major difference between California and the Northeastern ISOs is the separate PX in California. In New York, PJM, and (soon) New England, the short-term forward energy markets (day- and hour-ahead) are integrated within the ISO and its functions. In California, the PX operated, until January 2001, day-ahead and day-of (hourly) energy markets, while the ISO operates day- and hour-ahead markets for ancillary services as well as a real-time market for energy.

ISOs also differ in real-time dispatch. How far into the future (e.g., five minutes vs an hour) should the RTO forecast loads and generation and, therefore, its imbalance requirements? Who bears the risks of inaccurate forecasts? Looking ahead only one interval minimizes the risks of load-forecast errors but it also ensures that slow-moving generators will not be called upon. Looking ahead an hour increases load-forecast errors but permits the RTO to call on slow-moving resources and those that are not online but can be started within an hour.

The Appendix provides additional detail on the markets and operations in New York and California. It also discusses some of the problems the ISOs have faced in implementing their intrahour markets.

COMPETITIVE-MARKET PRICES

DAY-AHEAD AND REAL-TIME PRICES

If markets are efficient and producers and consumers are risk-neutral, energy prices in day-ahead and real-time markets should be, on average, roughly the same. Table 2 shows, however, that day-ahead prices are higher than real-time prices in PJM and New York but not in California. In 1999, the day-ahead and real-time prices in California were, on average, the same. For the first five months of 2000 (before the California markets exploded), the day-ahead prices were 10% less than the real-time prices. ISO New England has not yet opened its day-ahead markets; it currently operates real-time markets only.

As expected, hourly prices are much more variable in real-time than in day-ahead markets for the ISOs, as shown by the higher values for real-time standard deviations.[*] Prices should be more volatile in real time than a day ahead because of the unexpected events that can occur in real time, including forced outages of generation and transmission equipment and sudden weather changes. Indeed, real-time prices are sometimes negative when demand is so low that generators are running at their low-operating limit (LOL) and might have to be turned off (which would subsequently require the costs of a unit startup).

In PJM, the day-ahead and real-time prices are modestly correlated, with a correlation coefficient of 0.65.[#] For both northern and southern California, the correlation coefficients are roughly the same, 0.66 and 0.68, respectively. In New York, on the other hand, the two sets of prices are only weakly correlated, with a correlation coefficient of 0.32.

Day-ahead prices might be higher than real-time prices because of two factors. First, consumers who want to protect themselves from the higher volatility in real-time markets might be willing to pay more for energy in the day-ahead market for insurance. Similarly, suppliers worried about possible forced outages might want to sell more of their output in real time (at which point they know whether the unit is on or off); suppliers might therefore withhold some capacity from the day-ahead market to provide insurance against such outages. Both factors

[*]The standard deviation is a statistical term that measures the amount of variation among the values of a sample or population. It is equal to the sum of the squares of the differences between the individual values and the mean divided by one less than the sample size: $\sqrt{\sum(x_i - x_{mean})^2/(n-1)}$.

[#]A correlation coefficient of 0.65 means that 42% (square of coefficient) of the variation in one variable (e.g., day-ahead price) can be explained by the variation in the second variable (e.g., real-time price).

argue for higher day-ahead prices. However, the existence of some suppliers and some consumers who are not risk averse suggests that any difference between day-ahead and real-time prices will be arbitraged by these risk-neutral (or risk-seeking) market participants.

On the other hand, the real-time prices implicitly include generator maneuverability (ramprate) as well as energy because the units participating in this market are responding to 5- or 10-minute dispatch signals, in essence providing the load-following ancillary service. Not all generators (e.g., nuclear and large coal units) have sufficient flexibility to participate in intrahour markets; their relative inflexibility restricts them to participating in hourly markets. This rationale suggests that real-time prices should be higher than day-ahead prices.

Table 2. Statistical characteristics of day-ahead and real-time hourly energy markets[a]

	PJM 6/00 to 10/00		New York 1/00 to 10/00		California NP15 1/00 to 10/00		California SP15 1/00 to 10/00	
	DA	RT	DA	RT	DA	RT	DA	RT
Energy prices ($/MWh)								
Average	28.6	27.5	38.1	34.1	72.9	94.0	73.9	77.3
Maximum	140	597	523	838	1100	750	750	750
Minimum	0	0	0	-862	6	-326	0	-329
Standard deviation	18.5	21.8	23.0	37.7	78.1	108.6	88.4	107.4
Correlation between day-ahead and real-time prices	0.65		0.32		0.66		0.68	
Fraction of hours (%)								
< $10/MWh	4.2	3.3	0.2	4.2	0.9	8.3	1.5	14.5
> $200/MWh	0.0	0.1	0.2	0.5	4.4	14.8	7.3	9.5

[a]DA is day ahead and RT is real time. The PJM and New York data are for the control areas as a whole; the data for California are shown separately for the two major congestion zones, north of Path 15 and south of Path 15. Prices were capped in PJM and New York at $1000/MWh; in the California ISO markets prices were capped at $750/MWh through June 30, 2000, at $500/MWh from July 1 through August 6, and at $250/MWh from August 7 on.

More generally, day-ahead and real-time prices will diverge when either suppliers or consumers are able to exercise market power or when market rules yield inefficient outcomes.

Two additional factors unique to the California markets might explain why real-time prices were higher than day-ahead prices (FERC 2000c). Because price caps were used in the ISO markets but not the PX markets, load-serving entities (primarily the distribution utilities) had strong incentives to shift demand from the PX day-ahead market to the ISO real-time

market when demand was high and prices were likely to be high. Suppliers also had strong incentives to shift generation to the ISO market for replacement reserve under such conditions because they could get paid twice, once for the capacity and again for the energy.[*] FERC (2000e) eliminated the double payment for generators to remove the economic incentive they had to wait for the real-time market to sell power. In addition, FERC imposed a penalty on loads that purchase more than 5% of their requirements from the ISO's real-time market. (FERC offered no rationale for the 5% limit.)

REAL-TIME HOURLY AND INTRAHOUR PRICES

Intrahour prices, by definition, are more variable than hourly prices (Table 3).[#] For one zone in PJM (PEPCO), intrahour prices vary by, on average, $10/MWh (see top of Fig. 3) and the trend in prices (a shift from increasing to decreasing or vice versa) changes about once every two hours. For PJM, intrahour price changes average more than one-third of the average hourly price ($10.4 vs $27.1/MWh).

Prices are even more volatile in New York and California. In California, intrahour prices vary by almost $50/MWh, and the price trend changes sign about 1.5 times an hour. The average price change from one 10-minute interval to the next is about $17/MWh, compared with $4 in New York and $1 in PJM. (New York and PJM use 5-minute intervals.)

As shown in Fig. 3, intrahour prices can vary substantially, especially for New York. These large price changes from interval to interval suggest the importance of examining closely the operations of these balancing markets. Even for PJM, the intrahour price changes are substantial and merit attention. For example, intrahour price changes within PJM are greater than $5/MWh for about 60% of the hours.

Figure 4 shows intrahour prices for 12 hours on one day for the PJM PEPCO zone. For several hours, from 2 to 5 am, prices were stable, both within each hour and from hour to hour. However, from 6 am through noon, the intrahour price changes were substantial. Between 6 and 7 am, for example, the price increased from $17 to $30/MWh, then decreased to $25/MWh, and then increased again to $35/MWh.

---

[*]Under the $750 cap, in place through June 2000, suppliers could receive as much as $1500/MWh, $750/MW-hr of capacity for replacement reserve plus $750/MWh for energy supplied from that capacity. Beginning in early August, the ISO capped the replacement-reserve price at $100/MW, limiting the payment to no more than $350/MWh ($100 for the capacity plus $250 for the energy).

[#]The hourly price in these real-time ISO markets is equal to the weighted average of the intrahour interval prices, where the weights are the amounts of energy bought or sold each interval.

Table 3. Statistical characteristics of intrahour energy markets for September and October 2000

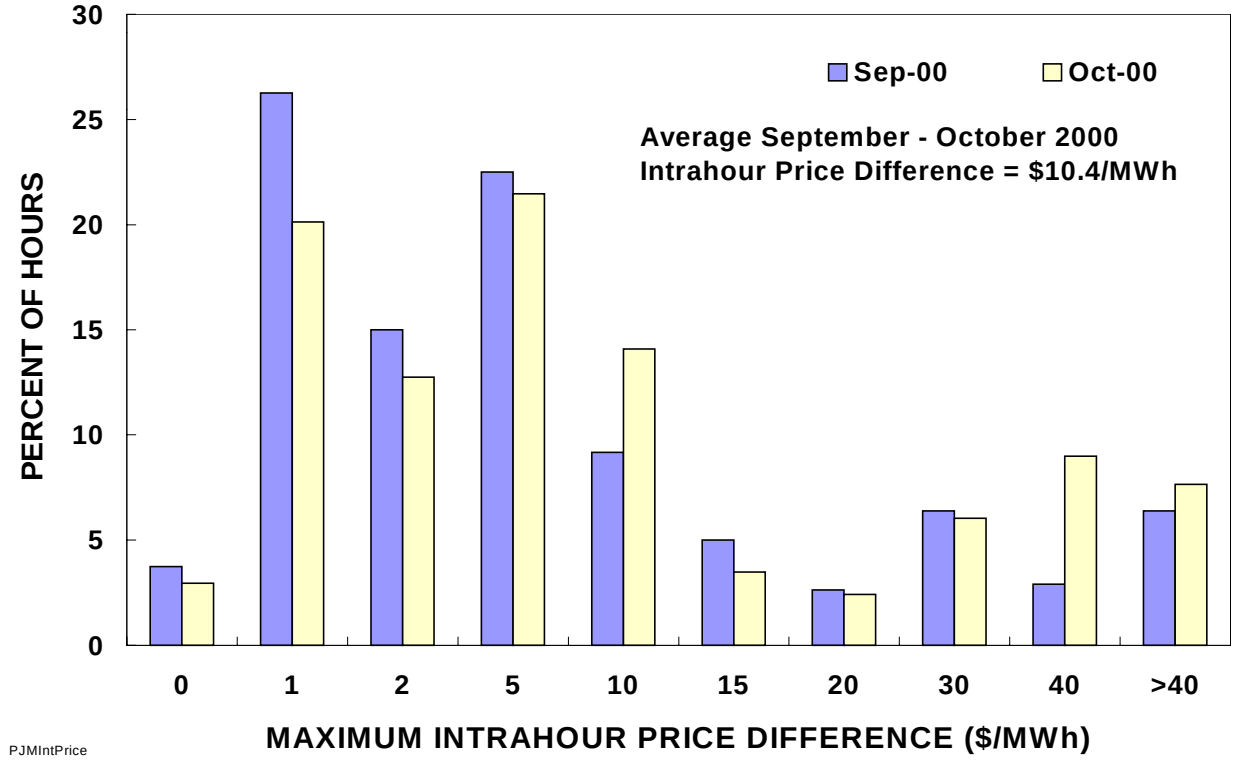
	PJM - PEPCO	New York - West	California - NP15	California - SP15
Energy prices ($/MWh)				
Average	27.1	38.3	152.6	90.8
Maximum	175	223	250	250
Minimum	0	-627	-75	-250
Standard deviation	20.7	36.8	78.3	84.4
Average interval-to-interval price change[a]	1.2	3.6	18.9	16.1
Average intrahour price change[b]	10.4	22.4	48.7	47.0
Number of sign changes per hour	0.5	4.4	1.7	1.4

[a]This statistic is the average of the absolute value of the difference between the price during one interval and the preceding interval, $|P_t - P_{t-1}|$.

[b]This price change is the average of the differences between the highest and lowest interval prices within each hour.

SUMMARY

These data on day-ahead, real-time, and interval prices show considerable volatility. This volatility is in large part a consequence of the dynamics of electricity supply, demand, and costs. The volatility is also a function of problems in market design and implementation, some of which are discussed in the Appendix.



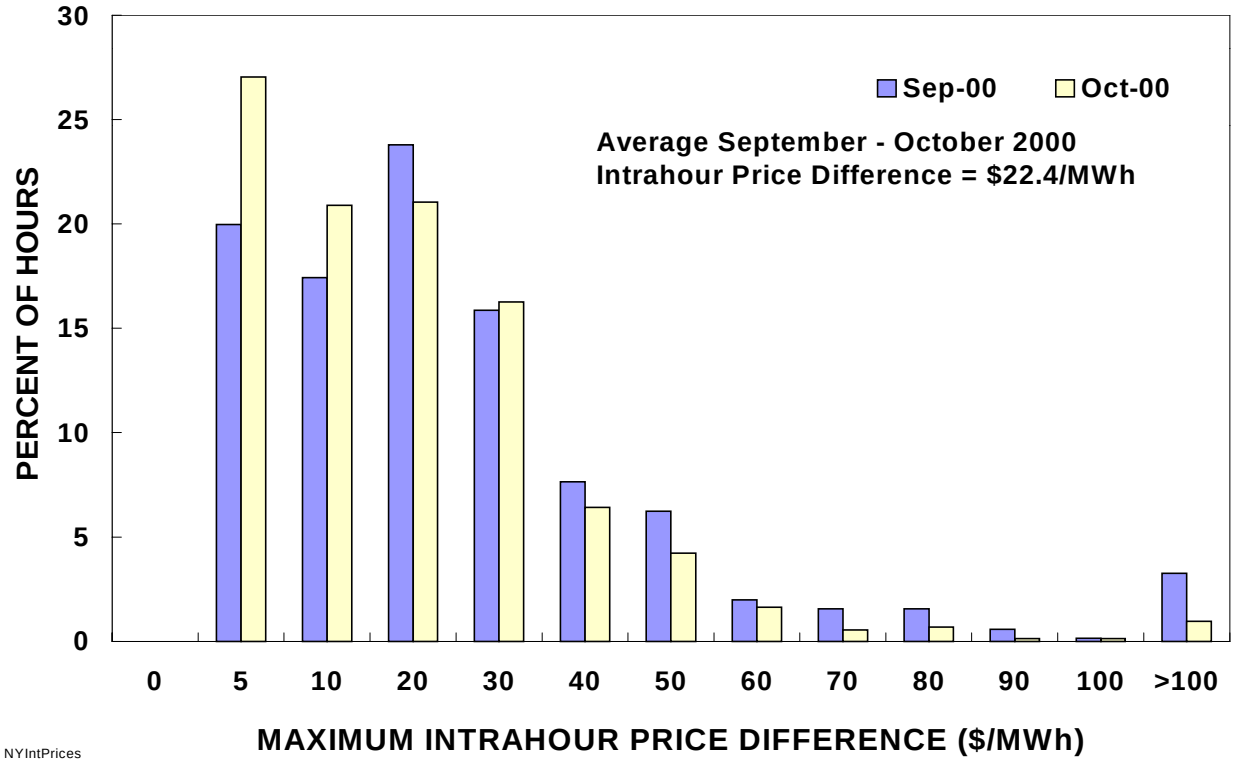


Fig. 3. **Intrahour energy price differences in PJM's PEPCO zone (top) and New York's West zone (bottom) for September and October 2000. Note the different scales for the *X* axes of the two graphs.**

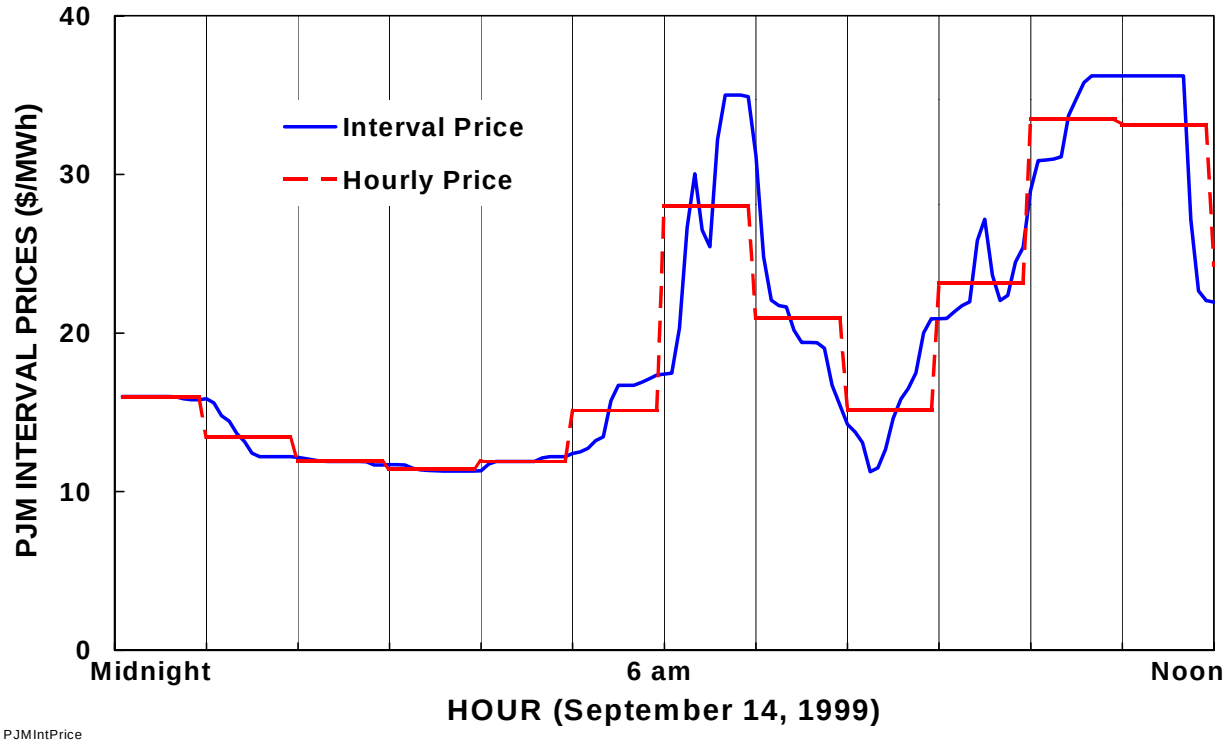


Fig. 4. **Intrahour (interval) and hourly electricity prices for the PJM PEPCO zone for 12 hours on one day. PJM calculates and posts prices every five minutes.**

EXAMPLES

INTRODUCTION

This chapter presents several hypothetical examples to illustrate the complexities in real-time operations and how they sometimes make determination of the MCP a difficult task.

Generation and load schedules are never completely accurate. Real-time loads fluctuate above or below their schedules. Some generators will be unable to match exactly their schedules. (These unscheduled variations in generation and load are in addition to the random minute-to-minute fluctuations that are compensated for by the regulation ancillary service; the variations considered here are longer-term than those the regulation service matches.) To rebalance load and generation and "true up" the schedules to reality, it is necessary to have real-time balancing operations. And, if generation is separated from system control and generation is competitive, the system operator must run a real-time market with prices set every interval.

In its simplest form, running a real-time balancing market involves only two steps. First, decide how much more or less generation will be needed in the next interval. Second, select the least expensive generators that can meet the expected need. Selecting the appropriate units depends on the incremental and decremental energy bids submitted by those generators wishing to participate in the RTO's balancing market. Typically, these bids include several price:quantity pairs; each pair includes an incremental or decremental quantity (in MW) and the associated price (in $/MWh).

In general, when additional energy is needed, the least expensive unconstrained unit online at any time should set the MCP, the price all generators producing power at that time receive for their output. Conversely, when less energy is required, the most expensive unconstrained unit should set the MCP.

For example, a 500-MW unit scheduled to operate at 400 MW during the coming hour might submit two incremental bids: the first 50 MW at $25/MWh and the second 50 MW at $30/MWh. This unit might also submit decremental bids, the prices it is willing to pay the RTO to reduce its output and, instead, purchase that energy from the system. For this unit, the decremental bids might be $23/MWh for the first 50-MW decrement and $20/MWh for an additional 25-MW decrement. In their simplest form, these bids include only the incremental costs or savings to produce more or less electricity. These bids would be dominated by the generator's fuel cost and incremental heat rate. The bids can change abruptly, however, when the generator is near a discontinuity in its operating range. If increasing output requires turning

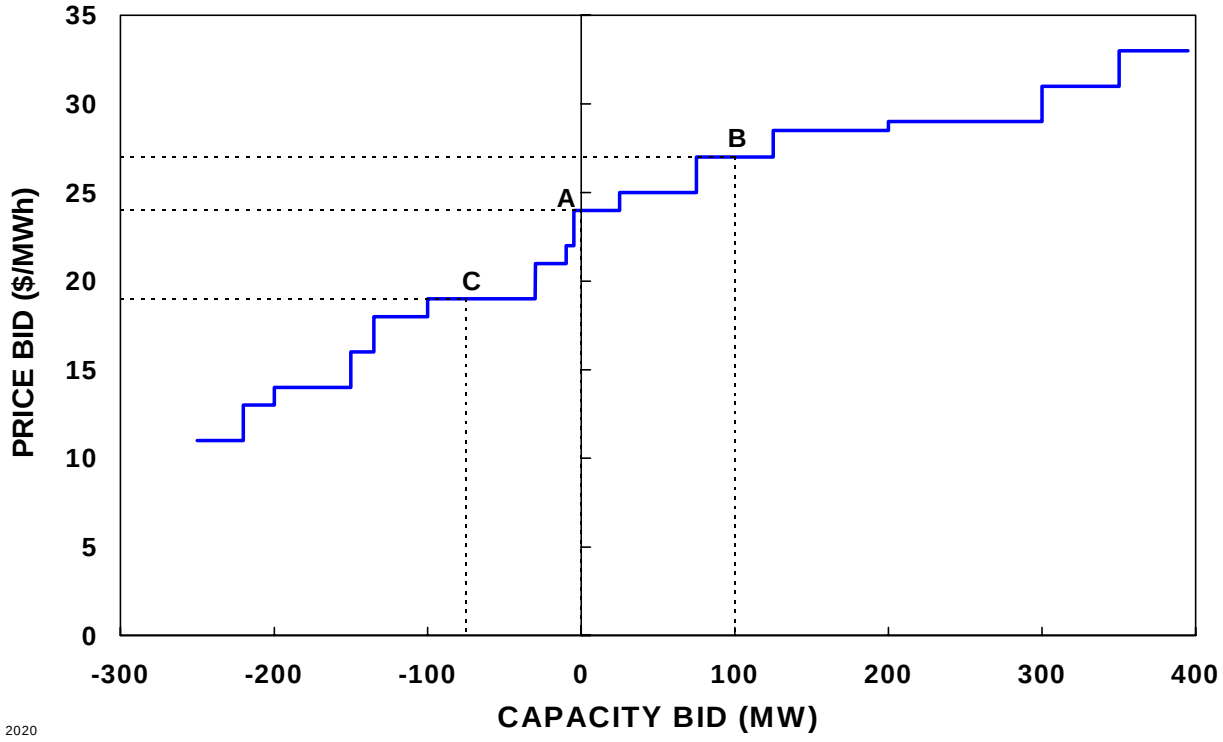


Fig. 5. **Stack of generator bids in a real-time balancing market. The zero point on the horizontal axis (A) represents the current operating point. Bids to the right are for increments and bids to the left are for decrements. If the load is expected to increase by 100 MW, the RTO would acquire resources up to point B at a price of $27/MWh. However, if the load is expected to drop 75 MW, the RTO would decrement resources to C at a price of $19/MWh.**

on additional coal mills, pumps, fans, or other auxiliary equipment, for example, the owner may bid the next increment of output at a higher rate than the fuel cost and heat rate alone would indicate.

To aid in selecting the appropriate resources, the RTO will first "stack" the incremental bids in order of increasing bid price and stack the decremental bids in order of decreasing bid price.* Figure 5 illustrates the form such a resultant stack might take. In this example, the system is currently in balance at point A with an implied price of $24/MWh. If the RTO expects loads to grow or the output of generating units to decline by 100 MW, it will move up the stack until it has dispatched an additional 100 MW of generation, to point B. Specifically, it buys 25 MW from the supplier that bid $24, 50 MW from the supplier that bid $25, and 25 MW of the 50 MW offered at $27. The MCP is then $27/MWh, and all three suppliers receive this price

*If some of the decremental bids are higher than some of the incremental bids, the RTO should clear them right away. In other words, when some market participants are willing to pay more to buy power on the spot market than others are willing to sell it for, those trades should be made immediately. The California ISO rules prohibit it from making such economically efficient trades.

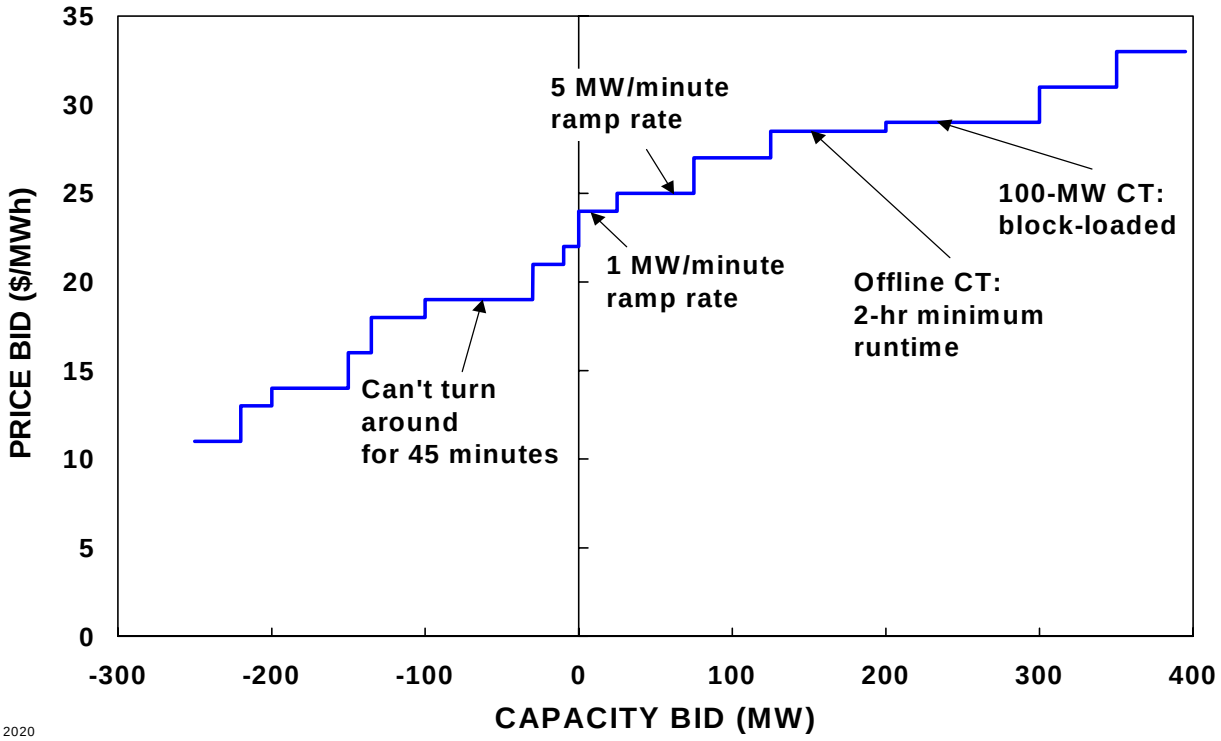


Fig. 6. **Stack of generator bids in a real-time balancing market. Unlike the bids in Fig. 5, these include various constraints that limit the units' flexibility in following RTO dispatch instructions.**

for the imbalance energy provided during this interval. Alternatively, the RTO might expect load to decline by 75 MW, in which case it would sell decremental energy down to point C. The unit agreeing to buy up to 70 MW of energy at a price of $19/MWh would set the MCP during this interval.

If generators were infinitely flexible (i.e., had no intertemporal constraints), the balancing problem would be solved, and there would be little need for this report. However, generators have many constraints that prevent them from turning on and off immediately and moving from one output level to another without delay. Figure 6 schematically illustrates some of these constraints, which complicate real-time operations and determination of the correct MCP. These constraints and their effects on MCP include:

- Minimum and maximum loadings, the range outside of which the generator cannot operate. A generator operating at its maximum output, for example, cannot be permitted to set the MCP when additional generation is required.

- Ramprate, the speed (in MW/minute) with which a generator can move from one output level to another. A unit that is decreasing its output at its maximum rate cannot be permitted to set the MCP when less generation is needed.

- Acceleration, the rate (in MW/minute^2) with which a generator can change its ramprate. Although control centers currently do not consider acceleration explicitly, it determines the speed with which a generator can change direction, say from moving up to moving down.

- Startup time, the amount of time it takes to bring a generator from an offline condition to the point that it is generating electricity at its minimum output level. The startup time depends on the amount of time the unit has been offline. The longer the offline time, the longer the startup time because the unit has gotten colder.

- Startup cost, the amount of money (primarily for fuel, but also for labor) to bring a unit from a cold condition to one in which it is ready to produce electricity.

- Minimum runtime, the minimum amount of time a unit must remain online and producing electricity. A unit that is within its minimum runtime should not be permitted to set the MCP unless the RTO needs the unit that hour and it is the most expensive unit then online.

- Block loading, a generator that can run at only one or a few discrete output points. This situation typically occurs with CTs that can be operated only at full output. Such a unit should not be permitted to set the MCP unless no other units have been backed down to accommodate the output of this unit.

These and other physical and financial constraints complicate the dispatch and short-term commitment decisions the RTO must make to maintain the necessary generation/load balance. The first three factors (upper and lower limits, ramprate, and acceleration rate) affect unit-commitment and dispatch decisions. The last four factors (startup time and cost, minimum runtime, and block loading) affect unit-commitment decisions; but once the unit is turned on, these factors do not affect dispatch. The following sections develop several examples to illustrate how these constraints affect dispatch and MCP.

RAMPRATE LIMIT

Consider the situation shown by the squares in Fig. 7, in which three units are on the margin at an MCP of $21/MWh. The chart shows the incremental-fuel-cost curves for three generators whose outputs are not limited by any of the constraints listed above. In competitive electricity markets, the units would bid their incremental and decremental outputs in real-time energy markets at prices based on these short-run marginal costs. In this example, the three units together produce 654 MW (289 MW for Unit 1, 175 MW for Unit 2, and 190 MW for Unit 3) at a market price of $21/MWh.

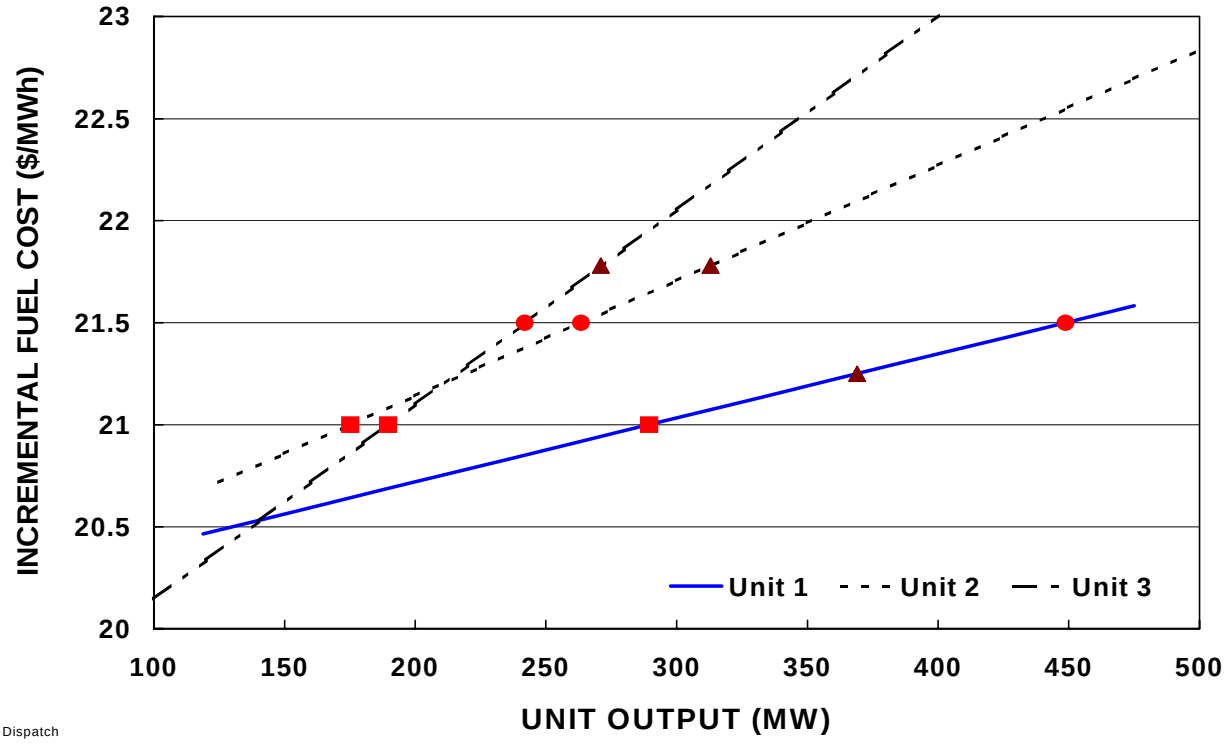


Fig. 7. **Incremental-cost curves for three generators. The squares show the outputs of the units when the MCP is $21/MWh. The circles and triangles show the outputs under two conditions when an additional 300 MW is needed: (1) all three units are unconstrained (circles) and (2) Unit 1 can move only half as far as optimal (triangles).**

What if the RTO needs an additional 300 MW? Based on the bids from the three units (assumed here equal to their marginal-cost curves), the outputs from the three units would increase by 160 MW for Unit 1 (to 449 MW), by 89 MW for Unit 2 (to 264 MW), and by 52 MW for Unit 3 (to 242 MW). The increments differ across the three units because of the slopes of the marginal-cost curves. In this case, the outputs of the three units increase from the points represented by the three squares at $21/MWh to the three circles at $21.5/MWh, as shown in Fig. 7.

What would the dispatch be, however, if Unit 1 had a low ramprate and was able to move only half as far as required (from 289 to 369 MW, instead of to 449 MW) during this interval? In this case, units 2 and 3 would need to increase outputs more (to 313 and 271 MW, respectively). The resultant dispatch, shown by the triangles in Fig. 7, has Unit 1 operating at a cost of $21.25/MWh and units 2 and 3 operating at $21.78/MWh. What, in this case, is the MCP? Unit 1 cannot set the MCP because it is constrained (specifically, it is moving as fast as it can and is unable to increase output beyond the point represented by the triangle). Therefore, units 2 and 3, both of which are unconstrained, set the MCP at $21.78/MWh. In this case, the ramprate limit of Unit 1 raises the MCP from $21.5 to $21.78/MWh.

This example shows that energy cost (a function of heat rate and fuel cost) is only one factor that affects real-time prices. As loads change from hour to hour, and especially as generation and load imbalances occur within an hour, the speed with which generator output can change is important and can affect both operations and market prices.

LOW-OPERATING LIMIT

Consider the situation represented by the circles in Fig. 8, in which the three units discussed above are operating at the same marginal cost of $21.5/MWh. If the RTO decides it needs additional generation and the amount required (270 MW in this case) is more than the three units can provide, it may need to turn on a new unit. This new unit almost certainly will have a higher energy bid than the three units already operating. If this new unit is small in size (with a minimum load *less* than 270 MW), its output can be accommodated without backing down any of the cheaper units already online. In such a case, the new unit, because it is unconstrained, will set the MCP. However, if this new unit is sufficiently large (with a minimum load of *more* than 270 MW), the RTO may operate it at its LOL.

To accommodate the additional output from the unit at LOL (420 MW in this example), the outputs from the three marginal units must be reduced, in this example by a total of 150 MW as shown by the squares in Fig. 8. In this case, starting a new, more expensive unit and running it at its LOL increases production costs but lowers the MCP from $21.5/MWh to $21.25/MWh. The MCP is set by the three unconstrained units rather than the unit at LOL because the LOL unit is constrained on. This reduction in MCP cuts generator earnings. In addition, the cost to start and operate the new unit run at LOL will likely be collected from electricity consumers through an uplift charge.

Although running a unit at LOL appears to be undesirable from the perspective of efficient markets, RTOs might often find themselves in such a situation. If the RTO's short-term forecast shows a need for more generation than can be provided by the units then online, it might have to turn on a new unit to meet expected conditions. Given the RTO's reliability responsibilities, it may sometimes find itself in situations where market-based decisions conflict with reliability decisions. A market-driven RTO might let the interval prices increase in the expectation (hope?) that the high prices will encourage generation owners to supply more output in real time. Alternatively, to ensure the necessary generation/load balance, the RTO must pay to start up a new generator, even if doing so artificially suppresses market prices. The magnitude of these costs is a function of the size and ramprate of the units that need to be turned on. If such units are small in size and have high ramprates, they may be able to set the MCP. Large, slow moving units, however, may require out-of-market decisions that suppress the MCP and result in uplift charges.

Consider a slightly more complicated situation than the one discussed above. Assume the RTO needs an additional 100 MW during the next interval. The two marginal units together can only provide an additional 80 MW (60 MW from Unit 1 and 20 MW from Unit 2). The

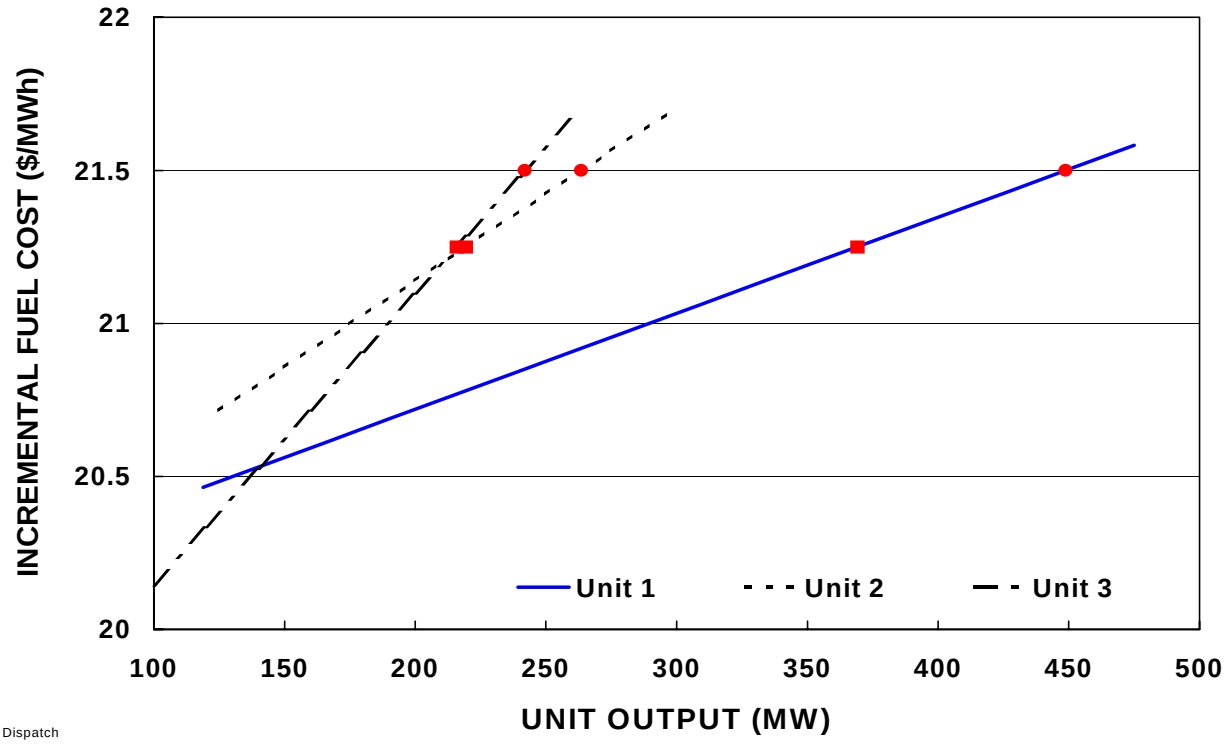


Fig. 8. **The circles at $21.5/MWh show the initial outputs of three generators. Turning on a new generator and operating it at its LOL requires the outputs from these three generators to be cut by 150 MW, yielding an MCP of $21.25/MWh (squares).**

RTO turns on a new unit that has an LOL of 40 MW. The remaining 60 MW are provided least expensively by increasing output from Unit 1 by 40 MW and Unit 2 by its remaining 20 MW. Because Unit 1 is the only unconstrained unit, its bid at its new output level sets the MCP, equal to $22.3/MWh, as shown in Fig. 9.

The Federal Energy Regulatory Commission (2000d) recently ruled on just such an issue involving the price of energy when lower-cost units are dispatched down to accommodate the start of a more expensive unit. In this case involving the New York ISO, it "… determined that the least expensive unit to be backed down, not the fixed-block resource, will set the market-clearing price."

STARTUP COSTS

If the RTO's forecasts show a need for additional generation, it may decide to start one or more units currently offline. Deciding which units to start depends on its expectations for future generation requirements and their duration. Assume the RTO is deciding between two CTs. CT1 has a startup cost of $40/MW and an energy bid price of $40/MWh. CT2 has a startup cost of $15/MW and an energy bid price of $50/MWh. If the unit is expected to operate for only an hour or two, CT2 is the preferred choice (Fig. 10). However, if the RTO expects to

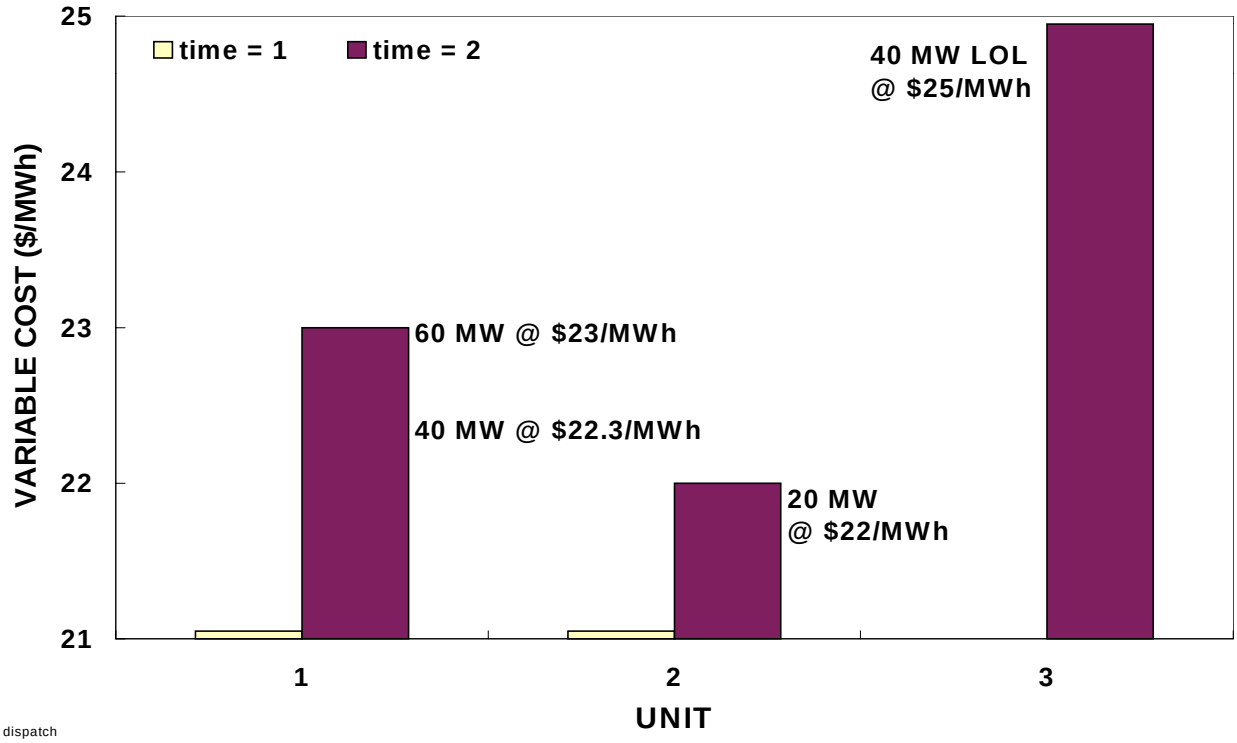


Fig. 9. **At time 1, units 1 and 2 are on the margin with an MCP of $21/MWh. To meet the increased demand of 100 MW at time 2, the RTO runs a new unit at its LOL of 40 MW, dispatches the remaining 20 MW from Unit 2, and dispatches 40 of the remaining 60 MW from Unit 1. Because Unit 1 is the only unconstrained unit, it sets the MCP at time 2 at $22.3/MWh.**

run the unit for more than two hours, CT1 is cheaper. The RTO's decision affects uplift (the amount customers pay for startup costs) and may also affect the MCP if the CT is the marginal unit. The RTO decisions can also dramatically affect the profits of CT1 and CT2.

This example illustrates a general problem with real-time operations and markets: the time period to use in forecasting future requirements and who bears the risks of such forecast errors. This issue is especially important for CTs and other resources that can start up within a few minutes and have minimum runtimes on the order of an hour. If high spot prices last for only a few 5-minute periods, it is not worthwhile to start and run such units. However, if high prices persist for several intervals, it may be profitable to start and run such units. Should the RTO issue short-term price forecasts for the next several intervals? Should generators be at risk for these unit-commitment decisions? Should the RTO decide when to start these units and then guarantee that, at a minimum, they will recover their startup and fuel costs and will operate for at least their minimum runtimes?

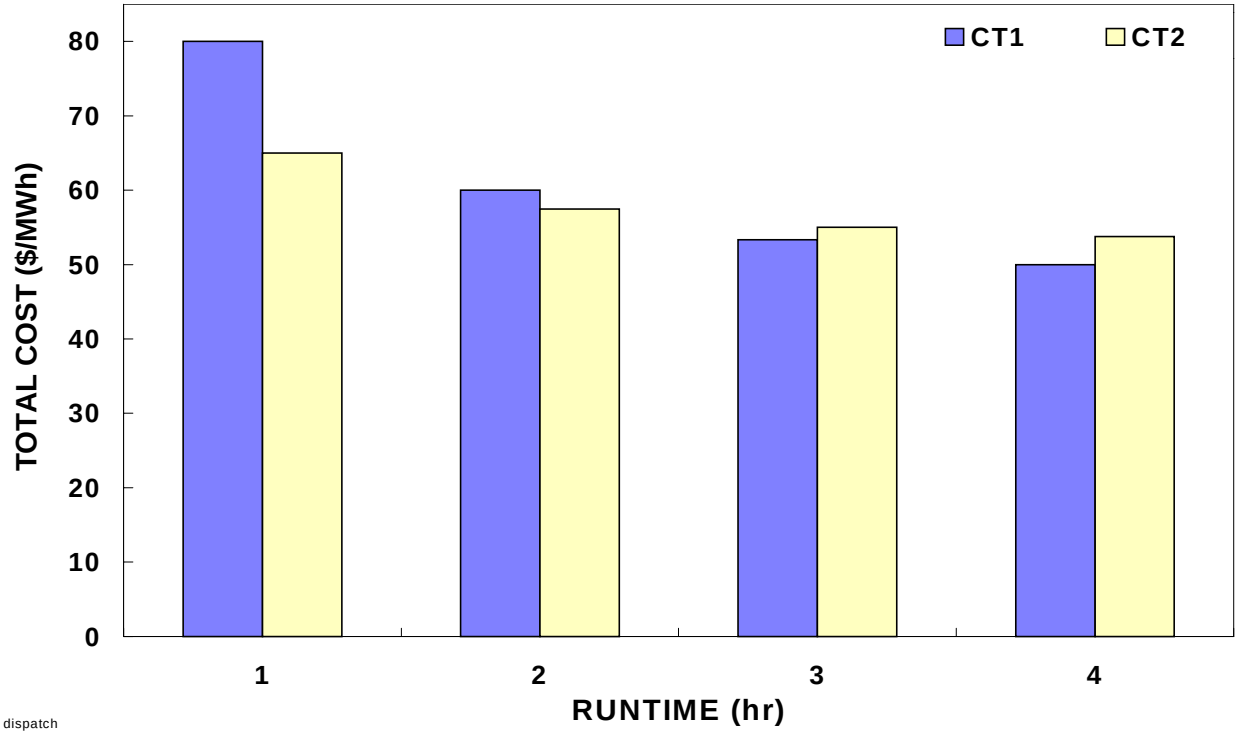


Fig. 10. Total cost per MWh to start and operate CTs as a function of the expected runtime and energy bid price.

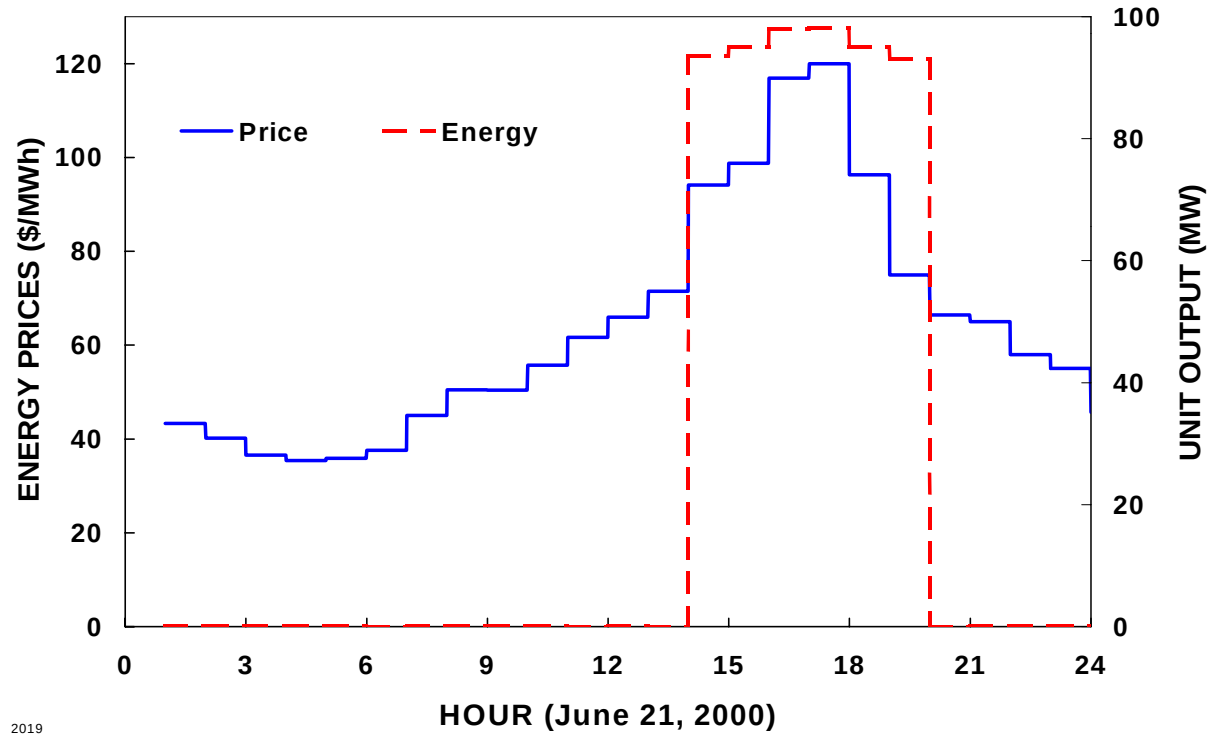
ENERGY-LIMITED UNITS

Some generators, in particular hydroelectric units, have limited amounts of fuel (impounded water in this case) they can use during any given time period. Thus, unlike fossil and nuclear units, which are capacity constrained, these units are energy constrained. (Fossil units with fixed emissions allowances might also be treated as energy-constrained units.) The challenge for these units is to optimize their limited energy output to maximize earnings. Solving this problem is challenging for at least two reasons.

- The generator owner (or RTO) must forecast future energy prices and then schedule the unit on the basis of the forecast.

- The optimal schedule requires a tradeoff between output and efficiency (bottom of Fig. 11).

In this case, the revenue-maximizing solution is to run the unit at close to maximum output during hours 16 and 17, when prices exceed $100/MWh, and at slightly lower levels during four other hours, as shown in the top of Fig. 11. Running the unit at maximum output would yield 5% less electricity overall and cut revenues by almost 3%. Running the unit at maximum efficiency would yield 3% more electricity and cut revenues by 1%.



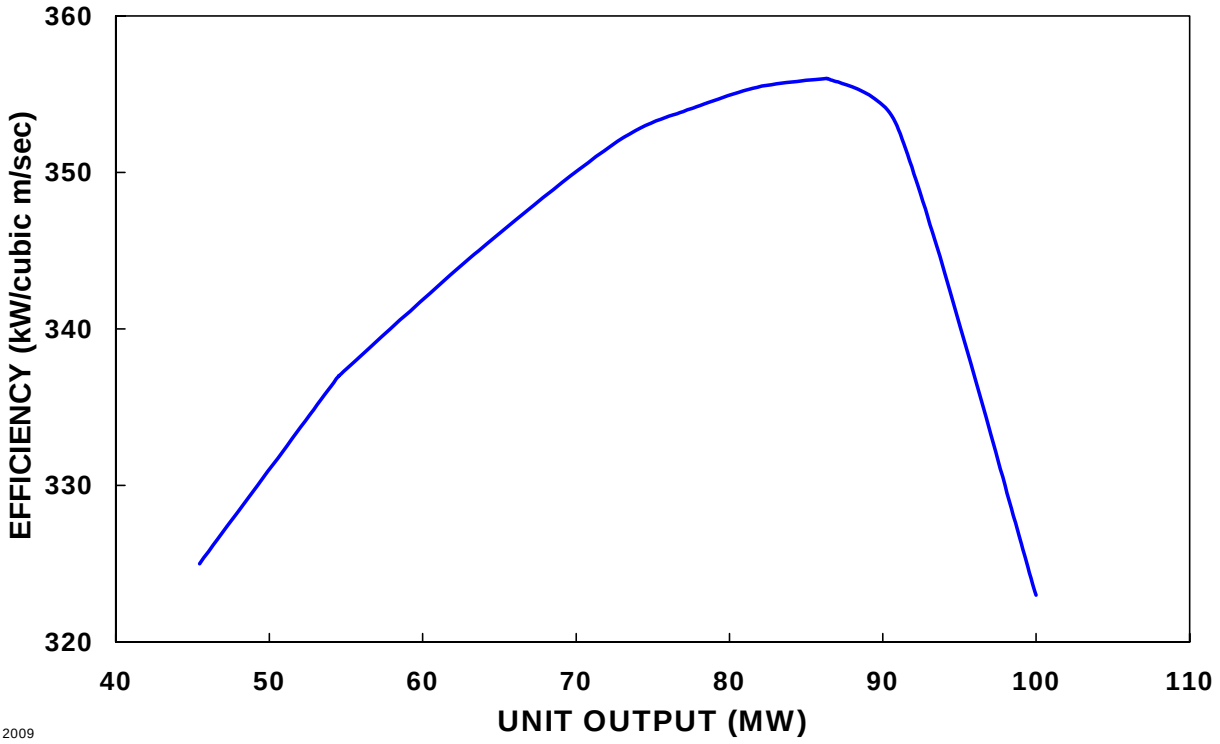


Fig. 11. **An energy-limited hydro unit will generate when prices are highest (top). Complications concern (1) the tradeoff between efficiency (the amount of electricity produced per cubic meter of water) and unit output (bottom) and (2) uncertainty about real-time energy prices.**

The unit-commitment and dispatch decisions for a pumped-storage hydro plant are similar to those discussed above, but more complicated. Pumped-storage units consume electricity during low-price hours to pump water into a reservoir; they then produce energy during high-price hours by running water from the reservoir through their turbines. This pump-generate cycle typically operates every day. Deciding when and how much to pump and when and how much to generate depends on the plant's pumping and generation characteristics (capacity, ramprates, turnaround time, efficiency, and storage capacity) as well as expected energy prices throughout the day.

RAMPRATE LIMITS AND MULTIPLE INTERVALS

An earlier example in this chapter (Fig. 7) examined the effects on dispatch and MCP of units with different ramprates. Here I extend that example to consider multiple intervals. In this case, the RTO foresees a need for an additional 100 MW; in addition, this 100-MW increment is expected to be needed for several 5-minute intervals.

Unit 1, with a bid of $26/MWh can fully meet this 100-MW need. However, its ramprate (8 MW/minute) is such that it can provide only 40 MW more each 5-minute interval. Unit 2, on the other hand, can provide the full 100 MW almost immediately, but its bid price is $28/MWh. Given this pair of bids, the RTO would take the maximum available from Unit 1 in interval 1 (40 MW) and 60 MW from Unit 2 (Fig. 12). Because Unit 1 is ramping at its maximum rate, the MCP during interval 1 is set by Unit 2 at $28. In going from interval 1 to 2, the RTO would increase the output from Unit 1 by its maximum, an additional 40 MW, and *reduce* the output from Unit 2 by the same amount to meet the unchanged 100-MW requirement. Once again, Unit 2 sets the MCP at $28/MWh. In going from interval 2 to 3, the RTO calls for an additional 20 MW from Unit 1 and dispatches Unit 2 to zero (because Unit 1 now provides all of the required 100 MW). In this final interval, Unit 1 sets the MCP at $26/MWh. The prices for the three intervals are $28, $28, and $26/MWh even though the balancing need is unchanged during the three intervals.

This example illustrates how different units can set the MCP at different times (Unit 2 in intervals 1 and 2 and Unit 1 in interval 3). It also shows the value of a high ramprate, which permits Unit 2 to temporarily raise the MCP. Finally, it shows how an RTO can simultaneously request some generators to move up and other generators to move down. Determining the MCP under such circumstances can be difficult.

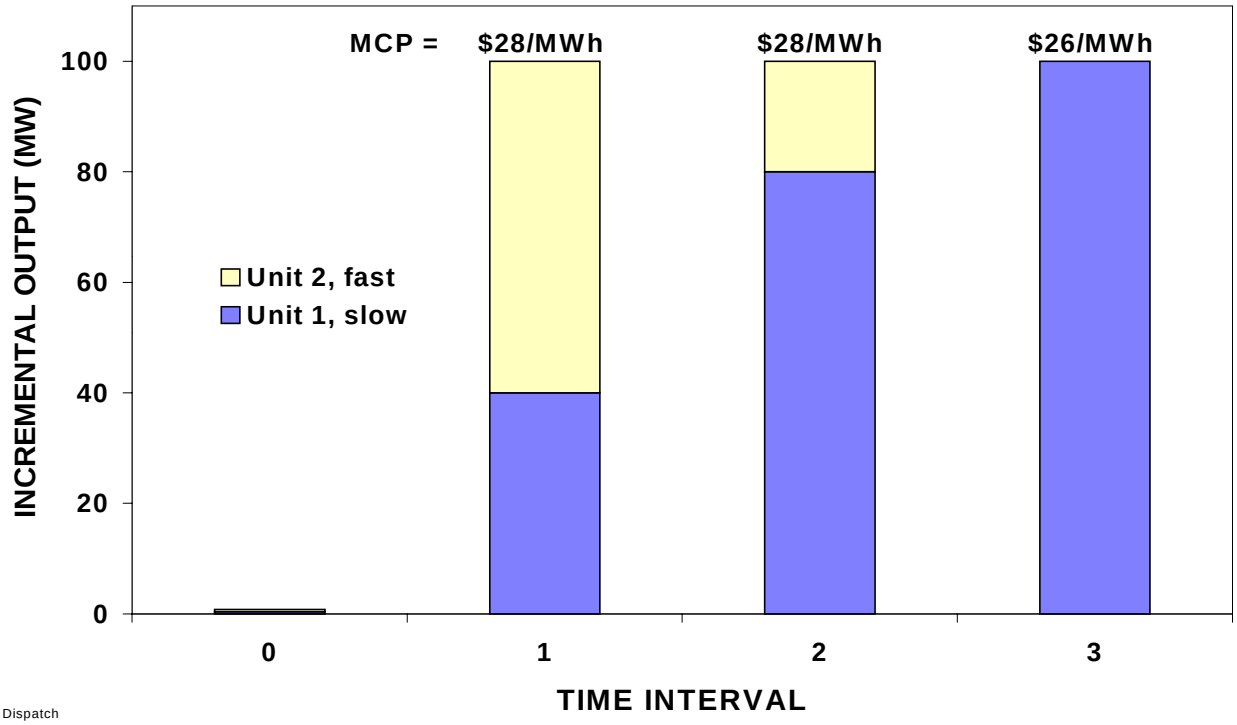


Fig. 12. **Outputs from units 1 and 2 during three time intervals when the total incremental generation needed is 100 MW. Unit 1 can ramp at no more than 40 MW each interval and has a bid of $26/MWh, while Unit 2 can ramp at twice that rate and has a bid of $28/MWh.**

IMPORTS AND EXPORTS

System operators treat generators inside their control area differently from those located in other control areas. These differences are a consequence of the interchange scheduling conventions among North American control areas, which are set by NERC's Policy 3 – Interchange. Typically, control areas schedule power transfers from one to another (called interchange) only at the top of each hour; that is, schedules are rarely adjusted during the hour.* (The system operator can dispatch generators inside its control area at any time.) These schedule changes occur over a 20-minute period in the west and a 10-minute period in the east. When the change in imports from one hour to the next is large relative to changes in system load, the RTO might need to dispatch some internal generation in the opposite direction to maintain generation/load balance, as illustrated in Fig. 13.

A large schedule change from one hour to the next is likely to impose large imbalances that switch signs at the top of the hour because of differences in the load and schedule

*PJM permits interchange schedules to change every 15 minutes with a maximum net interchange of 500 MW during each period.

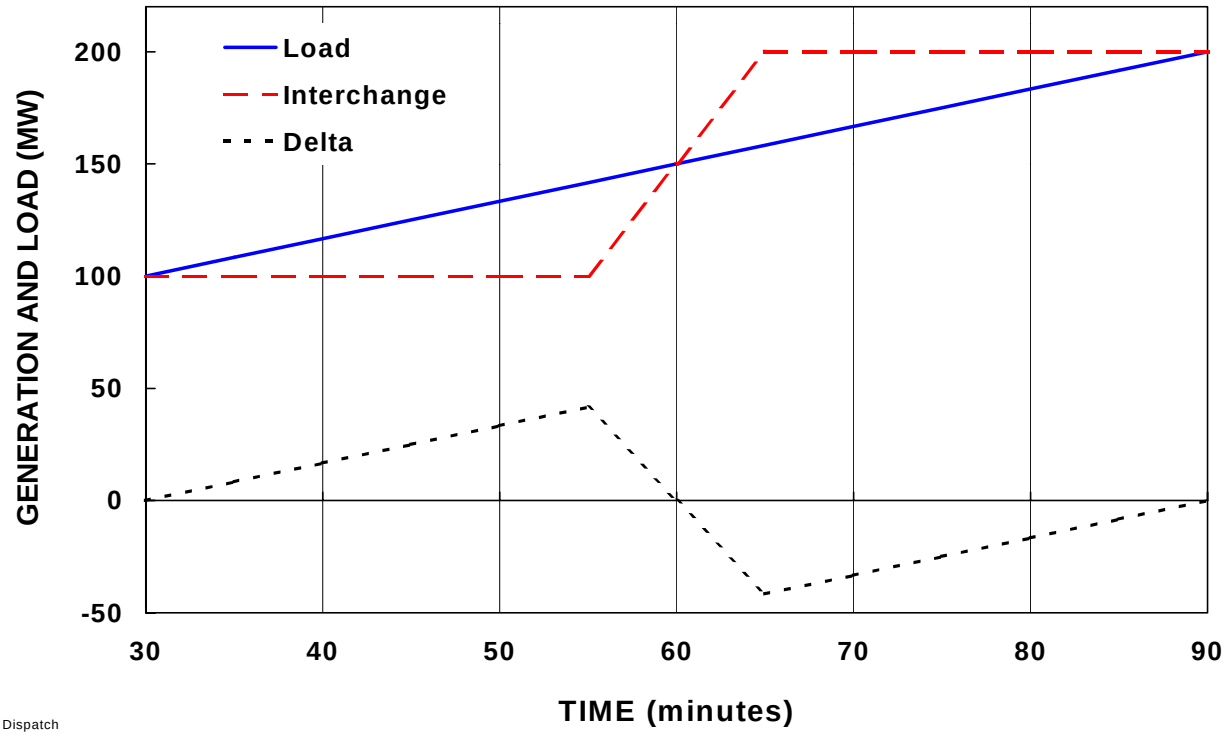


Fig. 13. **Increase in imports to a control area from 100 MW in one hour to 200 MW in the next hour (dashed line) and the associated change in the load to be served by this schedule (solid line). The bottom part of the figure (dotted line) shows the change in imbalance during this 60-minute period.**

ramprates. Specifically, the 10-minute top-of-the-hour schedule change is six times the hourly change in load. Thus, an increase in schedule from one hour to the next causes a generation deficit during the second half of the first hour and a generation surplus during the first half of the subsequent hour. Thus, the RTO must increment increasing amounts of generation between minutes 30 and 55 of the first hour (presumably increasing the MCP), then rapidly decrement these units during the next ten minutes (decreasing the MCP), and then, once again, incrementing generation and increasing the MCP for the next 25 minutes. Large schedule changes are most likely to occur during the morning rampup and evening dropoff.

Interchange schedules may create other complications related to their once-an-hour limitation. If an external generator bids into the real-time energy market as a dispatchable resource, the RTO might accept it for the first interval of the hour. However, for the remainder of the hour (e.g., the remaining eleven 5-minute intervals), the unit is no longer dispatchable because changing its output level would require coordination between the operators of the two control areas. Who is responsible for any out-of-market costs associated with the continued operation of this unit during intervals when its bid price is higher than the MCP? Should the generator be at risk if, after the first interval, its bid price will be above the MCP? Or should the RTO guarantee the generator that, once dispatched by the RTO, its costs will be recovered through an uplift charge, if necessary, for the remainder of the hour? In neither case should

such a resource be permitted to set the MCP after the first interval of the hour. Also, in either case, the suppliers will adjust their bidding strategies to reduce risk and increase profits.

INTERVAL VS HOURLY SETTLEMENTS

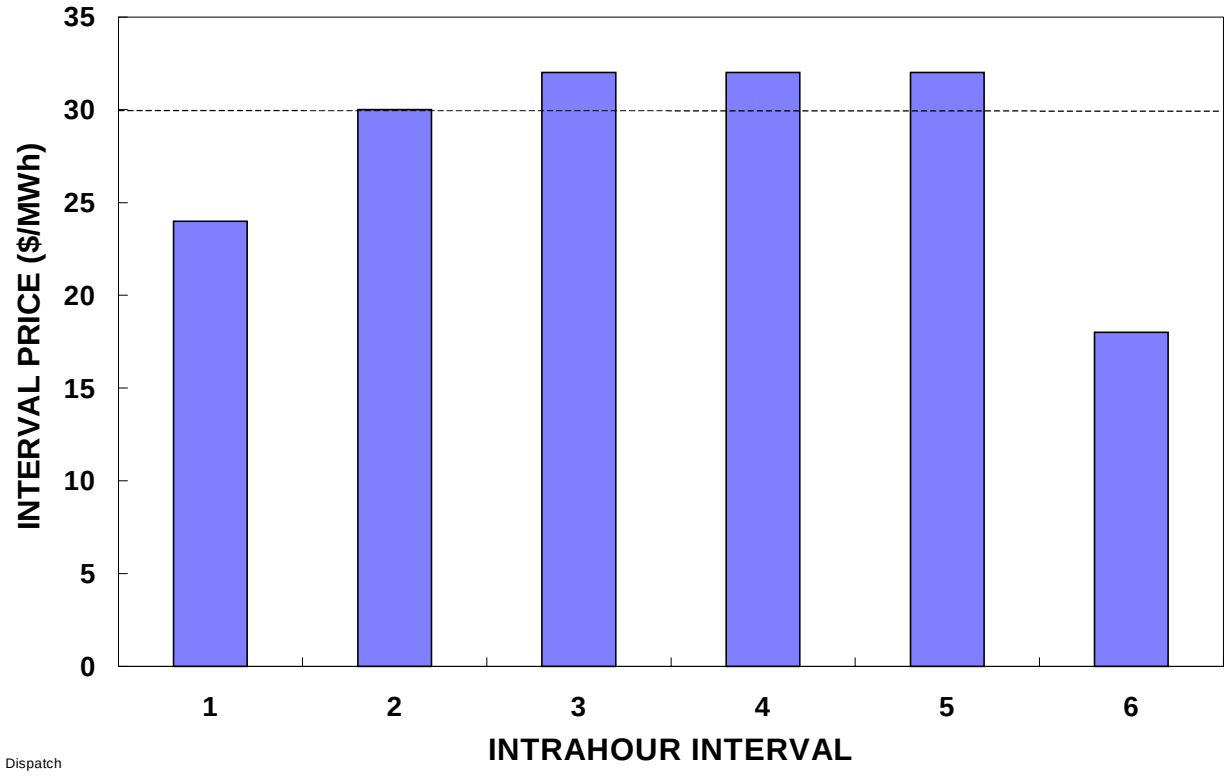
The three Northeastern ISOs currently use settlement systems that calculate prices for every 5-minute interval but perform financial settlements on the basis of hourly quantities and prices. The hourly price is calculated as the weighted average of the 12 interval prices with the weights equal to the quantities bought or sold during each interval. ISO New England (2000) recently installed an electronic dispatch system to automate the flow of ISO instructions to individual generators "to assure that compensation for generators is based on their actual dispatch and that proper incentives exist to follow ISO dispatch instructions."

Depending on the sequence of prices that occur during an hour and the maneuverability of the generators then online, the generators may have an incentive and the ability to arbitrage differences between interval and hourly prices. Because of such problems, the California ISO (2000a) switched from hourly to interval settlements in September 2000 in which both 10-minute prices and quantities are used to pay (or charge) generators for their output.

Consider the example shown in the top of Fig. 14, in which the price is $24 during the first interval, $30 during the next interval, $32 during the next three intervals, and $18 during the final interval of this hour. The hourly average price is $30/MWh.

A unit with marginal operating costs that vary between $20 and $28/MWh (as it moves from its minimum of 50 MW to its maximum of 150 MW) should operate at its optimal point of 100 MW during the first interval. Assuming this unit can ramp from one output level to another quickly, it would operate at its maximum output of 150 MW during intervals two through five when the price exceeded the $28/MWh cost of this unit at its maximum output. The sharp drop from $32 to $18/MWh between intervals 5 and 6 reflect the change in imbalance energy the RTO requires, from a positive to a negative requirement. The optimal output for this generator is 50 MW, its minimum level, assuming it will likely operate profitably again during the following hour. If settlements occur on an interval basis, this sequence of operations would yield earnings for this generator of $617 for the hour.

If, however, settlements are based on hourly quantities, the operator of this generator can, during interval 5, guess that the average price for the hour will almost surely exceed its operating costs. In that case, it would ignore the lower price during the sixth interval and continue to operate at full output, yielding earnings of $761 for the hour. (Ignoring the interval prices in this fashion would yield earnings of only $500 with interval settlements.)



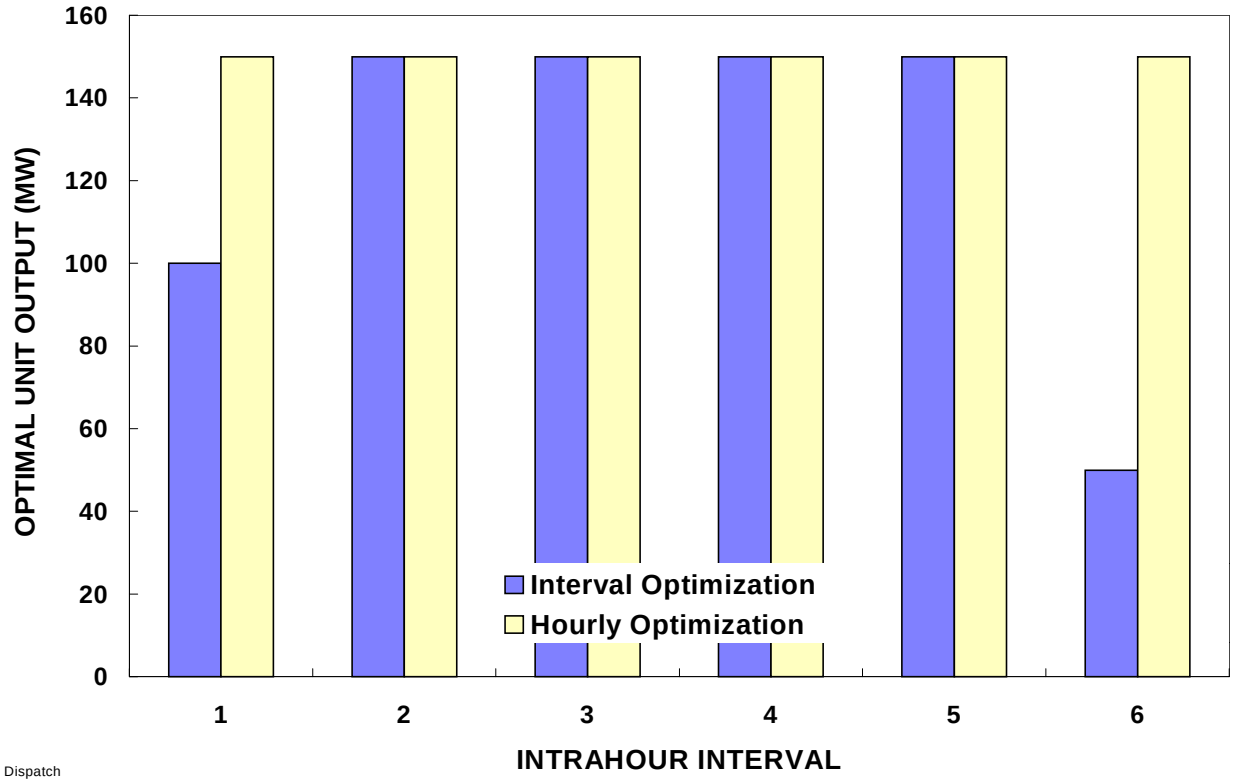


Fig. 14. **Ten-minute interval prices for one hour and the hourly average price (top) and unit output if settled on interval vs hourly basis (bottom).**

The California ISO (2000a) recently switched from a system with 10-minute dispatch instructions and hourly financial settlements to one with 10-minute settlements. It estimated an annual savings of almost $200 million from this change in lower costs for regulation, better response to ISO dispatch instructions, and more efficient pricing. The ISO noted that the mismatch between dispatch and settlement intervals means that scheduling coordinators (SCs) "can satisfy the ISO instructions at any time during the hour (e.g., an instruction issued by the ISO for the delivery of Energy … in the first or second ten-minute interval of an hour may be satisfied by the delivery of Energy anytime during the remainder of the hour). Thus, a Scheduling Coordinator has little or no incentive to deliver instructed Imbalance Energy during the BEEP [Balancing Energy and Ex Post Pricing] interval for which the Energy was dispatched by the ISO."

INTERMITTENT RESOURCES

Wind and some other renewable resources are intermittent, which means that their output cannot be controlled in real time (i.e., they are not dispatchable). In addition, the outputs from these resources cannot be scheduled as accurately as those from traditional fossil, nuclear, and hydro resources. Because the power output from a wind farm depends on the wind speed, scheduling wind-power output depends on one's ability to accurately forecast wind speeds.

In this example, I consider a control area that dispatches resources every five minutes to maintain its ACE within the NERC-prescribed CPS bounds. When resources are dispatched up (to compensate for undergeneration), prices increase at an assumed rate (specified in $/MW of imbalance). As additional incremental resources are called upon, the price increases further. Similarly, when resources are dispatched down, prices are lower than the hourly average. In this hypothetical example, the imbalance energy ranges from +146 MW to −197 MW with an hourly average of 0 MW and a standard deviation of 65 MW. Because of these moves up and down the generation-supply curve, the 5-minute interval prices range from $66/MWh to −$19/MWh, with an hourly average of $30/MWh and a standard deviation of $16/MWh.[*]

I next add the time-varying output from a wind farm to this control area. The output of the wind farm has two effects on the control area. First, the inability to accurately forecast wind output means that the hourly values of wind output will differ somewhat from the scheduled values of wind output. Second, the variability of wind output means that the system operator will need to dispatch resources up and down intrahour to offset this wind variability. Thus, wind resources impose hourly and intrahour costs on a control area.

In practice, the control-area operator would modify its intrahour dispatch to maintain ACE within the same CPS limits as without the wind output. Because the automatic-generation-control algorithm in most control area energy-management systems is very complicated, I

[*]The statistical properties of this interval-price series are roughly consistent with the interval prices in New York and California (Chapter 3).

cannot readily simulate such a dispatch. Therefore, I analyze two boundary cases, the first of which unfairly favors wind and the second of which unfairly penalizes wind:

- The system operator ignores the effects of the intermittent output from the wind farm and dispatches generation resources exactly as it did in the base case. This case favors wind because it exempts wind from any balancing costs.

- The system operator compensates fully for variations in wind output. That is, it dispatches other resources in exactly the same amounts and the opposite direction from wind. This case unfairly penalizes wind by requiring it to maintain a *perfect* energy balance at all times, unlike other resources which, in *aggregate* (not *individually*), are required only to maintain an *adequate* balance.

Although this is a hypothetical example, it uses 5-minute data from a wind farm. I selected a set of hours during which the wind output averaged close to 100 MW, with a range from 2 to 174 MW and a standard deviation of 20 MW. Because of the statistical nature of this example (time-varying dispatch and time-varying wind output), I ran this case 24 times. The results presented here are averages over these 24 hourly runs.

The changes in revenues received by this wind resource on a 5-minute basis, relative to those it would receive under hourly average prices, reflect two factors: the intrahour volatility of the wind output and the hourly difference between its actual and scheduled outputs. Operators of wind farms can do little about the first factor but they may be able to limit the second factor by developing accurate hour-ahead forecasts of wind output.

The top of Fig. 15 shows the percentage difference in hourly payments for the wind output relative to the hourly average price for the first case described above. If the system operator ignores the volatility of wind output, there are no costs associated with these intrahour variations in wind output. Thus, these results include the effects of scheduling error only (because there are no effects of volatility). The three curves represent positive, zero, and negative values of control-area imbalance offset. A positive offset means the system operator needs, on average, incremental resources during the hour to offset what would otherwise be an undergeneration situation. If the wind output is scheduled accurately (100 MW), the wind resource gets paid exactly what it would get paid on the basis of the hourly energy price. If the control area is, on average, energy deficient (i.e., the system operator must dispatch additional resources during the hour and the imbalance offset is positive), prices are higher, on average, than the hourly value. In such cases, wind gets paid more if it underschedules (i.e., produces more than its scheduled amount during the hour) and vice versa. For the cases analyzed here, the effects of varying imbalance offset and the accuracy of the wind schedule are less than 2% of wind revenues.

The bottom of Fig. 15 shows the difference in hourly payments to the wind resource for the case in which the system operator compensates 100% for the time-varying output of the

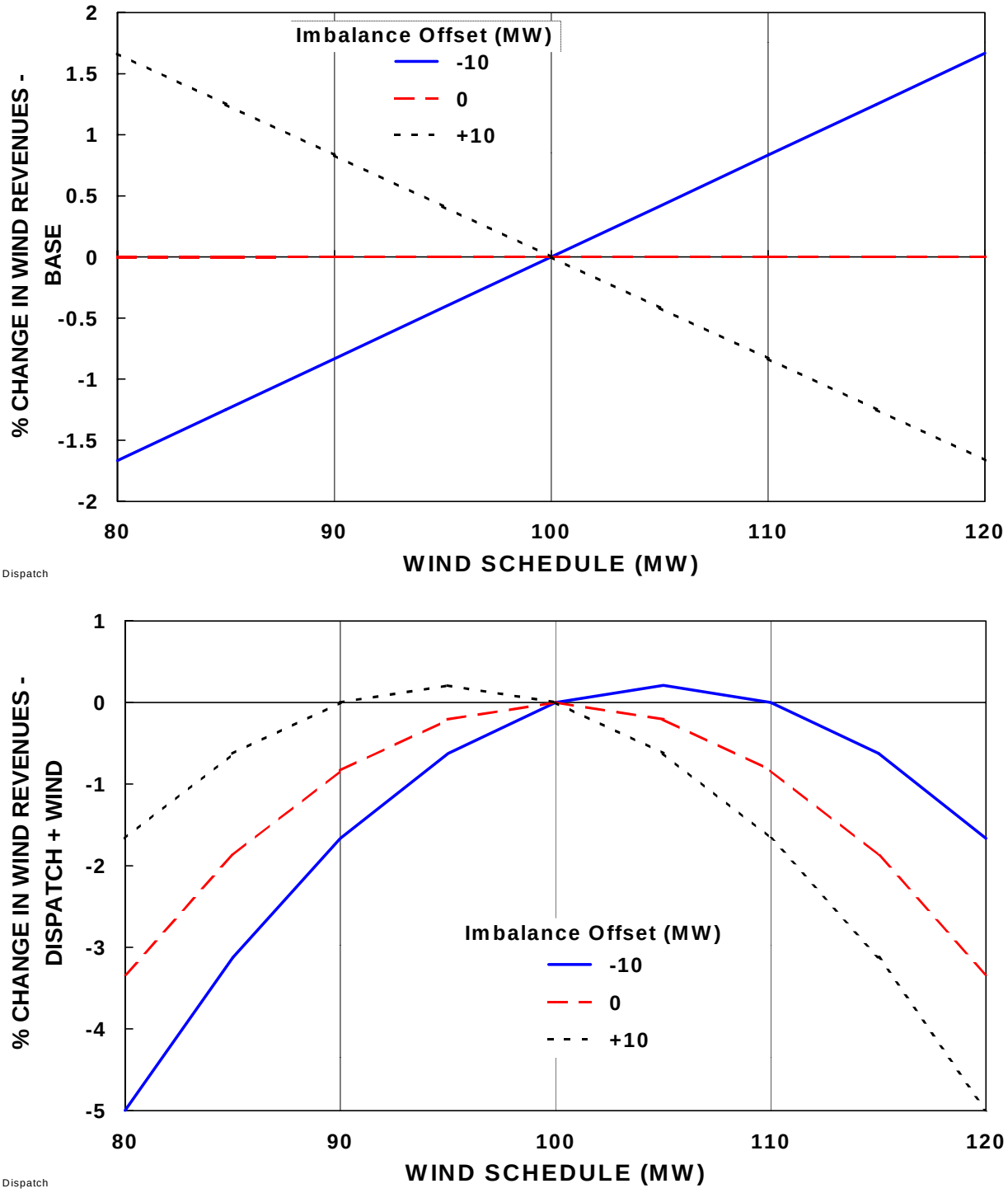


Fig. 15. **Percentage change in payments to a wind resource for energy as a function of control-area imbalance and wind schedule (vs actual wind output of 100 MW). A positive imbalance offset means the system operator needs, on average, incremental resources. The top graph shows results when the system operator ignores the energy-imbalance effects of wind. The bottom graph shows results when the system operator adjusts fully for the wind output.**

wind farm. The effects of intrahour variability in wind output are ignored to ease comparison with the top of Fig. 15. In this case, as expected, the wind resource collects less money than the hourly average price would imply regardless of the accuracy of its schedule and the control-area's overall hourly energy balance. The penalty imposed on wind in this case generally increases with increasing discrepancy between the scheduled and actual output of the wind farm. The compensation for the wind output is less for both under- and over-scheduling.

I also ran cases in which I doubled the size of the system imbalance to see how results depend on the size of the wind resource relative to the size of the imbalance market. Doubling the MW dispatch while holding the wind output unchanged reduces the dependence of wind revenues on its scheduling accuracy and the imbalance offset by almost half. It also reduces the penalty associated with the second form of ACE management (100% compensation for wind fluctuations). Thus, as expected, the smaller the wind resource relative to the size of the imbalance market (not the magnitude of the overall load), the fewer the adverse effects on its spot-market energy revenues. The size of the imbalance market is a function of the accuracy with which market participants meet their schedules, the ability of the system operator to manage ACE efficiently, the volatility of the load, and the size of the load.

Figure 16 shows the effects of intrahour wind volatility only on the hourly revenues a wind resource would receive, assuming it accurately scheduled its hourly average output. Under the first ACE-management case (in which the system operator ignores the effects of wind output on ACE), volatility has no effect on wind revenues and is, therefore, not shown on the figure. However, in the second case, in which the system operator compensates fully for the wind volatility, increasing wind volatility reduces the revenues received for the wind energy production. In this example, the wind farm would earn 3% less money than if it had no volatility. A wind farm with twice as much intrahour variation in its output would earn 13% less money.

Because this is a hypothetical example, one should not read too much into these results. However, the results point to a few overall conclusions. The amount of money wind resources will earn in real-time markets is a function of several factors, including their ability to schedule accurately (at least hour ahead); whether the control area is, overall, surplus or deficient (i.e., whether the system operator needs incremental or decremental generation); how the system operator redispatches resources in response to the time-varying wind output; the volatility of the wind output; and the correlation between the output of the wind resource and the control area's imbalance dispatch. If the control area is neither surplus nor deficient, if the wind output exactly matches the wind schedule, and if the wind output is entirely uncorrelated with the control area's dispatch, the wind resource will receive revenues that exactly match the hourly average price.

Under certain circumstances (e.g., the system operator ignores the volatility of wind output, the system needs additional resources that hour, and the wind resource generates more than its hourly schedule), the offset between actual and scheduled wind output can increase its

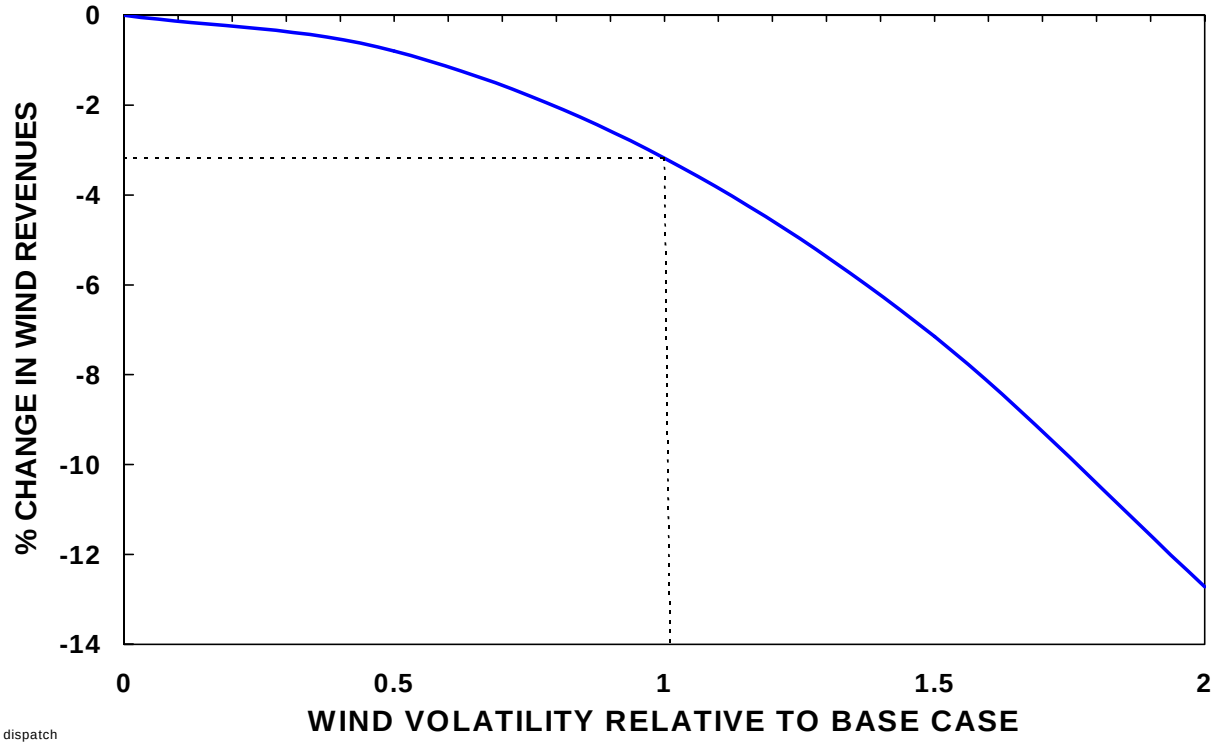


Fig. 16. Percentage change in wind-energy revenues as a function of the intrahour variability of wind output.

energy earnings. In general, however, wind will collect fewer dollars in the real-time imbalance market than it would if it could accurately schedule its output an hour ahead. If the numbers produced in this example are roughly correct, the volatility "penalty" for wind is on the order of several percentage points. The effects of scheduling error will depend on the relationship between the control area's need for resources and the sign of the wind scheduling error; in either case, the scheduling "penalty" will also be on the order of a few percentage points.

In New York, the ISO penalizes resources that overgenerate relative to their schedules by taking the "extra" energy without compensation. Resources that undergenerate are required to buy that energy from the spot market. The overgeneration penalty strikes me as unjustified because it has no cost basis. The undergeneration charge, on the other hand, is reasonable and consistent with the example developed here. In PJM, on the other hand, the wind resource would be a price taker, and its revenues would be computed as illustrated in this example. In approving 10-minute settlements in California, FERC (2000a) wrote "… as long as generators supplying energy on an uninstructed basis do not receive more for their energy than it is worth at the time it is supplied,… a punitive approach is neither necessary nor appropriate."

CONCLUSIONS

Real-time operations and pricing are essential to the proper functioning of competitive wholesale markets for three reasons. First, system operators must have enough generation and load resources available in real-time to maintain grid reliability; the most cost-effective way to obtain such resources is through the purchase and sale of energy every five or ten minutes. Second, real-time prices are the basis for all forward contracts, including long-term bilateral contracts, block monthly futures contracts, and day- and hour-ahead prices. Third, investment decisions are determined in part by expected real-time electricity prices.

Unfortunately, the design and operation of real-time markets are as complicated as they are important. Even though these real-time functions and operations are largely unchanged from those that exist with vertically integrated utilities, the separation of generation from system control and the need (or opportunity!) to use markets rather than command-and-control methods complicates their design. Ultimately, once the startup and transitional problems are worked out, these markets should provide a solid foundation for efficient and competitive wholesale electricity markets.

Four such markets operate today in North America, three in the Northeastern U.S. and one in California. The three Northeastern ISOs (probably because of their tight-power-pool history) chose a fundamentally different design than did California. California's approach is very decentralized, leaving many decisions to individual suppliers. As a consequence, the risks of poor decisionmaking fall primarily on the individual suppliers in California. The Northeastern ISOs, on the other hand, operate markets that are much more centralized. In particular, the Northeastern ISOs conduct a centralized day-ahead unit commitment that, to a large extent, removes the risk of poor decisionmaking from the individual suppliers and shifts those risks to retail customers in general.

Beyond these broad differences, the details of market design and operation are quite different among all four operational ISOs. In particular, the New England, New York, and California ISOs have experienced many startup problems. These problems have led to major outcries from market participants, especially the generators and power marketers, but sometimes also the load-serving entities. The ISOs have been working very hard to identify and resolve problems as they arise. For example, the California ISO has, as of December 2000, filed more than 30 tariff amendments with FERC to incorporate these changes. As PricewaterhouseCoopers (2000a) noted, "As with most other start-up organizations, management of ISO-NE has been challenged by the complexities of a rapidly changing organization and business environment. ISO-NE management has succeeded in establishing

effective practices for many of their operational responsibilities. However, significant room for improvement exists in most of the areas subject to this Operational Study."

Documentation of the design and operation details is either largely nonexistent or buried in obscure, hard-to-locate documents. As a consequence, it is difficult to draw conclusions on the lessons learned from the ISO experiences to date. For example, it appears that the PJM design and implementation are more efficient and less trouble-prone than the others, but I am unable to offer specific recommendations on what other ISOs and RTOs should do to improve their systems.

A brief review of the RTO filings with FERC shows a surprising neglect of the complexities of designing the intrahour balancing market that FERC requires of RTOs. The authors of these RTO filings were either unaware of the practices and problems in the existing ISOs or, more likely, did not have sufficient time to learn from these real-world experiences. As markets open in the Midwest, Texas, Ontario, Alberta, and other regions, problems analogous to those that occurred in the existing ISOs are likely to arise, unless the designers and operators of these soon-to-open markets make serious efforts to learn what went wrong and why.

The examples developed here, although simplified, offer useful guidance on how to treat generator constraints and how to determine market prices for each intrahour interval. An underlying theme in these examples is the use of markets (rather than direct control and penalties) to motivate behavior that maintains grid reliability. For example, the analysis of the effects of wind on grid operations shows costs imposed on the grid from an inability to (1) accurately schedule wind output ahead of time and (2) control output in real time. These costs appear as lower revenues to the wind owner than would occur if the resource could be accurately scheduled and then operated throughout the hour at its schedule. However, these lower revenues are not the consequence of an arbitrary penalty but, rather, reflect the actual costs the RTO experiences in buying and selling energy intrahour.

Ultimately, these real-time markets may work well, permitting system operators to replace command-and-control requirements with market signals (e.g., eliminating the need for contingency reserves). Such markets will likely lower electricity costs to consumers, efficiently deploy resources where they are most valuable, and guide investments in new generation.

ACKNOWLEDGMENTS

I thank Edison Electric Institute, Enron Wind Systems, and the Project for Sustainable FERC Energy Policy for their financial support of this project. I thank Terry Black, Kenneth Linder, Mathew Morey, and Russell Tucker for their advice and support throughout the course of this project. I thank my long-time colleague Brendan Kirby for his assistance and detailed review of an early draft of this report. I thank Terry Bilke, Terry Black, Tim Bush, James Caldwell, Mario DePillis, Benjamin Hobbs, William Hogan, Kenneth Linder, Mathew Morey, John Nielsen, Shmuel Oren, Larry Ruff, and Steven Stoft for their helpful comments on a draft of this report. Finally, I thank Fred O'Hara for editing the final report.

REFERENCES

Avista Corp. et al. 2000, *Supplemental Compliance Filing and Request for Declaratory Order Pursuant to Order 2000*, Docket No. RT01-35-000, submitted to the Federal Energy Regulatory Commission, Spokane, WA, October 16.

California Independent System Operator 2000a, *Amendment No. 29 to the ISO Tariff*, Docket No. ER00-2383-000, submitted to the Federal Energy Regulatory Commission, Folsom, CA, May 2.

California Independent System Operator 2000b, *Balancing Energy and Ex-Post Pricing*, Procedure No. M-403, Folsom, CA, July 14.

Carolina Power & Light Company et al. 2000, *Volume II, GridSouth Open Access Transmission Tariff*, Docket No. RT01-74-000, submitted to the Federal Energy Regulatory Commission, Raleigh, NC, October 16.

Competitive Market Group 2000, *ISO-New England Implementation and Operation of the Wholesale Electricity Market, Problems and Potential Solutions*, January 21; downloaded from ISO-New England website at www.iso-ne.com.

E. Hirst and B. Kirby 2000, *Measuring Generator Performance in Providing the Regulation and Load-Following Ancillary Services*, Oak Ridge, TN, December.

E. Hirst and B. Kirby 2001, *Retail-Load Participation in Competitive Wholesale Electricity Markets*, Edison Electric Institute, Washington, DC, and Project for Sustainable FERC Energy Policy, Alexandria, VA, January.

ISO New England, Inc. 2000, *Compliance Filing*, Docket No. EL00-014-000, submitted to the Federal Energy Regulatory Commission, Holyoke, MA, December 1.

Midwest Independent Transmission System Operator 2001, *Order No. 2000 Compliance Filing*, Docket No. RT01-87-000, submitted to the Federal Energy Regulatory Commission, Indianapolis, IN, January 16.

M. Morey 2001, *Power Market Auction Design: Rules and Lessons in Market-Based Control for the New Electricity Industry*, Edison Electric Institute, Washington, DC, forthcoming.

New York Department of Public Service 2000, *Interim Pricing Report on New York State's Independent System Operator*, Albany, NY, December.

New York Independent System Operator 1999a, *NYISO Transmission and Dispatching Operations Manual*, Schenectady, NY, September 1.

New York Independent System Operator 1999b, *NYISO Day Ahead Scheduling Manual*, Schenectady, NY, September 2.

New York State Electric & Gas Corp. 2000, *Complaint of New York State Electric & Gas Corporation to Suspend Market Based Rates for Energy Markets and Request for Emergency Technical Conference*, Docket No. EL00-70-000, submitted to the Federal Energy Regulatory Commission, Binghamton, NY, April 24.

North American Electric Reliability Council 1999, *NERC Operating Manual*, "Policy 1 — Generation Control and Performance," Princeton, NJ, November.

PJM Interconnection 2000a, *PJM Manual for Pre-Scheduling Operations*, Revision: 03, Norristown, PA, June 1.

PJM Interconnection 2000b, *PJM Manual for Dispatching Operations*, Revision: 06, Norristown, PA, June 1.

PJM Interconnection 2000c, *PJM Manual for Scheduling Operations*, Revision: 14, Norristown, PA, August 24.

PricewaterhouseCoopers 2000a, *ISO New England Inc. Operational Study: Final Report*, ISO New England, Holyoke, MA, June 7.

PricewaterhouseCoopers 2000b, *2000 Operational Study*, California Independent System Operator, Folsom, CA, November 9.

L. E. Ruff 2000, *RTOs, ISOs, Transcos and Transmission Pricing*, n/e/r/a, San Francisco, CA, April 29.

San Diego Gas & Electric 2000, *Answer of San Diego Gas & Electric Company in Support of Joint Motion for Emergency Relief and Further Proceedings*, Docket Nos. EL00-95-000 et al., submitted to the Federal Energy Regulatory Commission, San Diego, CA, October 19.

U.S. Federal Energy Regulatory Commission 1999, *Regional Transmission Organizations*, Order No. 2000, Docket No. RM99-2-000, Washington, DC, December 20.

U.S. Federal Energy Regulatory Commission 2000a, *California Independent System Operator, Order Conditionally Accepting Proposed Tariff Revisions*, Docket No. ER00-2328-000, Washington, DC, June 29.

U.S. Federal Energy Regulatory Commission 2000b, *New York Independent System Operator, Inc., Order on Tariff Filing and Complaint*, Docket Nos. ER00-3038-000 et al., Washington, DC, July 26.

U.S. Federal Energy Regulatory Commission 2000c, *Staff Report to the Federal Energy Regulatory Commission on Western Markets and the Causes of the Summer 2000 Price Abnormalities, Part I of Staff Report on U.S. Bulk Power Markets*, Washington, DC, November 1.

U.S. Federal Energy Regulatory Commission 2000d, *New York Independent System Operator, Inc., Order Extending Bid Cap, Acting on Tariff Sheets, and Establishing Technical Conference*, Docket Nos. ER00-3591-000 et al., Washington, DC, November 8.

U.S. Federal Energy Regulatory Commission 2000e, *Order Directing Remedies for California Wholesale Electric Markets*, Docket Nos. EL00-95-000 et al., Washington, DC, December 15.

A. J. Wood and B. F. Wollenberg 1996, *Power Generation, Operation, and Control*, second edition, John Wiley & Sons, New York, NY.

APPENDIX: ISO EXPERIENCES AND RTO PLANS

The first two sections of this Appendix explain how the New York and California ISOs run their day-ahead and intrahour markets. The third section summarizes some of the complaints raised about the operation of these markets by various market participants. The fourth section offers specific examples of problems related to load forecasting and uplift charges. The final section summarizes some of the RTO plans for complying with the real-time balancing-market requirement for Order 2000.

NEW YORK

Figure 17 illustrates the scheduling and operations used by the New York ISO (1999a and b). The Bid/Post system allows market participants to post generator and load bids, as well as schedules for bilateral transactions. The system is also used by the ISO to post results from the day-ahead and real-time markets, as well as the advisory results of the hour-ahead Balancing-Market Evaluation (BME).

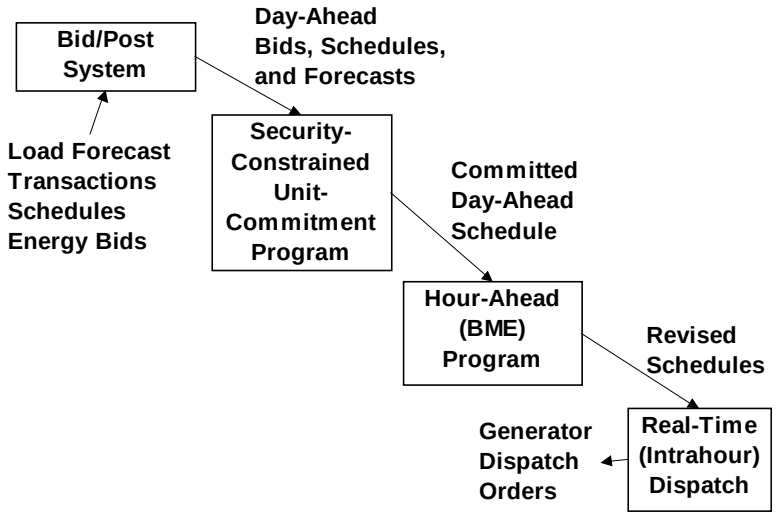


Fig. 17. New York ISO process for day-ahead scheduling and real-time operations.

The ISO's day-ahead scheduling process includes assembling data on planned transmission outages and the ISO's zonal load forecasts and using these data and forecasts as inputs to the ISO's security-constrained unit commitment (SCUC) model. The objective of the SCUC is to minimize the total bid production cost of meeting all purchasers' bids to buy energy a day ahead, provide enough ancillary services, commit enough capacity to meet the ISO's load forecast, and meet all bilateral transaction schedules submitted to the ISO.

The SCUC model is run several times to get a final solution for the following day's schedule of generation: (1) based solely on generator and load bids plus bilateral transactions and ancillary-service (operating reserve and regulation) requirements, (2) commitment of any additional generation needed if the ISO load forecast is higher than the market-participant

schedules, and (3) security-constrained dispatch to ensure that all the first-contingency requirements are met.

The SCUC considers several factors, including "current generating unit operating status, constraints on the minimum up and down time of the generators, generation and start up bid prices, plant-related startup and shutdown constraints, minimum and maximum generation constraints, generation and reserve requirements, maintenance and derating schedules, transmission constraints, phase angle regulator settings, and transaction bids" (New York ISO 1999b).

Generator inputs to the SCUC include up to 20 segments of an energy curve, as well as the noload cost ($/hr), startup bid ($), and startup and shutdown constraints (the number of times each day a unit can be started and stopped). The model adds transmission-loss penalties for each generator to account for system losses.

About 90 minutes before each hour, the BME unit-commitment model is run, with results (mostly, but not all, advisory) posted 30 minutes before the operating hour. Bids into the BME (used to determine the resources available for real-time dispatch) can include resources that are dispatchable within 5 minutes and can respond to dispatch requests plus fixed-block energy (nondispatchable) available for the next hour. In addition, the BME can consider modified and proposed new bilateral transactions, as well as bids to buy energy in the real-time market.

For real-time dispatch, the New York ISO classifies resources as on dispatch, off dispatch (but online), or offline but available.

According to the New York ISO (1999a), "The function of the SCD [security-constrained dispatch] program is to determine the least-cost dispatch of generation within the NYCA [New York Control Area] to meet its load and net interchange schedule, subject to generation, transmission, operating reserve, and regulation constraints. SCD performs this function nominally every five minutes as part of the real-time operation of the NY Power System." The SCD objective of minimizing cost is limited to the *incremental* bid cost of generation participating in the spot market.

The SCD treats CTs differently from steam units. It can consider startup or shutdown for CTs that have a shorter lead time than for steam units. In addition, CTs are, once turned on, dispatched at their full output (block loaded).

Just as the SCUC is run multiple times, the SCD is run twice, once for feasibility and the second time for optimization. The first run seeks a solution that meets load, the second run tries to improve on the first solution by finding a lower-cost combination of generators without violating any of the real-time security constraints. The New York SCD considers about 200 contingency states. Inputs to the analysis include telemetry values of generation output, power

flows on the transmission system, load, actual net interchange, and calculated losses. Additional SCD runs may be needed to turn CTs on or off.

Resources can set the 5-minute MCP only if they are not being dispatched against one of their limits. In addition, some resources outside the control area cannot set the market price because of constraints on hourly interchange schedules.

CALIFORNIA

The California ISO day-ahead process begins at 10 am, when SCs submit energy schedules and adjustment bids for congestion management to the ISO. (In California, the SCs must submit *balanced* schedules, which means that each SC must provide enough generation resources each hour to match its loads plus losses.) Between 10 and 11 am, the ISO analyzes the schedules and adjustment bids for any potential transmission congestion problems and, by 11 am, announces the results. If congestion is expected to occur, the SCs are given until noon to submit revised schedules and bids. At 1 pm, the ISO publishes the final day-ahead energy and ancillary-service schedules, the usage charges for all congested interfaces, and the final market-clearing prices for all ancillary services for each hour of the operating day.

Two hours before the start of the operating hour, SCs can submit updates to their energy and self-provided ancillary-service schedules, as well as new bids for the ISO hour-ahead ancillary-services auction. By one hour before the operating hour, the ISO hour-ahead ancillary service markets are complete. Within 45 minutes of the operating hour, SCs can submit supplemental energy bids for the ISO real-time market. During each operating hour, the ISO dispatches balancing energy and publishes energy prices every 10 minutes (compared to the 5-minute intervals used in the Northeast). The resources made available to the ISO for its real-time dispatch include incremental and decremental supplemental energy bids, as well as the operating reserve resources (which include spinning, nonspinning, and replacement reserves).

The California ISO operates two control rooms, one in its Folsom headquarters and the other in Alhambra. Altogether, 16 people work each shift, 10 in Folsom and 6 in Alhambra (PricewaterhouseCoopers 2000b, California ISO 2000b). These people and their functions include, in addition to the shift supervisor:

- Generation dispatchers (3) monitor the power system (e.g., ACE, automatic-generation-control, and reserves) and estimate and dispatch imbalance energy;
- Real-time grid resources coordinators (2) ensure that the resource stack in the 10-minute market is accurate, implement the 10-minute market, ensure that 10-minute prices reflect current dispatch, make any necessary out-of-market calls, and provide market information to the generation dispatchers;
- Real-time intertie schedulers (2) issue dispatch instructions for intertie schedule changes;

- Real-time schedulers (2) coordinate predispatched resources on interties with grid resources coordinators and notify SCs of dispatch instructions;
- Transmission dispatchers (4) monitor the California transmission system;
- Hour-ahead grid resources coordinator (1); and
- Day-ahead grid resources coordinator (1).

ISO EXPERIENCES

Experience to date with competitive real-time markets shows many problems. Three of the four ISOs now operating in the United States (New England, New York, and California, but apparently not PJM[*]) have experienced problems in the design and operation of their energy markets. Market participants in all three regions have complained about poor market design and implementation. This section summarizes some of the complaints raised about ISO operations; the following section gives two examples concerning load forecasting errors and energy uplift charges.

New York State Electric & Gas (NYSEG 2000) complained about many aspects of the New York ISO's (NYISO's) markets and their operation. As summarized by FERC (2000b), some of NYSEG's complaints are:

> Energy Price Fluctuations: NYSEG claims that the scope and frequency of the volatility of energy prices and the BME are difficult to comprehend, and that it has yet to receive a rational explanation from NYISO. NYSEG states that numerous extraordinary price spikes have occurred since the start of NYISO operations and that this volatility can render energy price prediction and planning to be of little or no value.
>
> Price Convergence: NYSEG states that NYISO market structure was based on the premise that the day-ahead and real-time markets would converge … . NYSEG conducted an analysis of the convergence of the day-ahead market prices and real-time market prices and concluded that the real-time market is consistently priced lower than the day-ahead market, and that the difference is significant ($2.88/MWh). NYSEG maintains that the price difference should not be this large or consistently in favor of the real-time market, and therefore concludes that the markets are not acting in a rational or competitive manner.
>
> Fixed Block Generation: NYSEG argues that NYISO's use of fixed block bidding in the real-time market creates … problems. NYSEG believes that

[*]It is not clear whether PJM has had fewer problems than the other ISOs because of the step-by-step deliberate approach taken by PJM to design and operation of its markets or because most of the generation in PJM is still owned and operated by vertically integrated utilities, PJM has ample generating capacity, and that capacity is relative flexible.

> NYISO has adopted a pricing rule not provided by the tariff, e.g., that the fixed block bid price is used to set the LBMP [locational based marginal price]. Therefore when NYISO dispatches a fixed block bid in the real-time market that is in excess of its needs, it must also back down the last most economic resource that was dispatched. The backed down resource must therefore "buy back" its dispatched energy at the LBMP set by the fixed block bid, thus the resource can be forced to purchase replacement energy in the real-time market at a price that exceeds its own bid price. NYSEG maintains that this pricing procedure is a violation of the tariff and the LBMP pricing theory. NYSEG states that generators that are dispatched down to make room for fixed block generators are being compensated using "Lost Opportunity Payments" (LOP) and that these LOP payments are not authorized for this purpose under NYISO services tariff. NYSEG maintains that the current NYISO method of letting the block bidding generator set the LBMP and compensating the dispatched down generator through an LOP … results in a higher LBMP, higher cost to consumers, and appears to not be authorized by the services tariff.

Appendix C of the FERC order (2000b) summarizes the ISO responses to each of NYSEG's complaints. The New York ISO explained its resolution of these and other problems at a January 2001 FERC technical conference; see www.nyiso.com/services/documents/filings/pdf/ferc_tech_conference/meeting_materials.html.

The Competitive Market Group (2000) complained about the number and complexity of the ISO New England market rules and the way the ISO administers its many markets. It claims, as an example, that:

> The clearing price for energy is often not set by the bid of the most expensive[*] [generator] being dispatched (the "marginal" unit). Although only the ISO can know precisely, it has been estimated that as much as half of the time, the marginal unit on the New England system is not the unit setting the clearing price. This is especially likely in periods of high demand and otherwise high clearing prices.
>
> The ISO has used a variety of means, including the use of "uplift" payments, "posturing," and "operator discretion" to dispatch units which are not considered in the calculation of clearing prices. Unless the market is allowed to function, which means the marginal unit on the system sets the clearing price, the aim of deregulation may never be realized.

---

[*] In theory and practice, the most expensive unit need not necessarily set the price; as noted correctly by the parenthetical phrase in this quote, the marginal unit(s) should set the price. As discussed in Chapter 4, units must be unconstrained by ramprate or load limits to be eligible to set the price.

San Diego Gas & Electric (2000), in its complaint about the California ISO, listed several "design and policy elements" that it felt needed correction:

- The ISO must operate short-run markets to maintain short-run reliability.
- Short forward energy and transmission/ancillary-services markets must not be separated (i.e., eliminate sequential and separate markets).
- Artificial restraints imposed on the ISO to separate markets should be removed, especially the requirement for balanced schedules.
- The ISO's short forward markets should serve physical bilateral schedules on a comparable basis with pool transactions. The ISO should be neutral in its transmission, ancillary services, and other requirements between bilateral and pool transactions.
- The ISO should offer (on a voluntary basis) least-cost dispatch and efficient congestion management.
- The ISO should offer a unit-commitment service on a voluntary basis.
- The ISO should resolve all congestion in each short forward market.
- The ISO should price locational (congestion) effects accurately, using nodal rather than zonal prices.
- The ISO should eliminate portfolio bidding and, instead, require unit-specific bidding.

In each of these cases (and many others like them), the ISO response typically agrees with some of the problems noted by the complainant, disagrees with some of these problems, and notes how hard the ISO is working to resolve these problems. In fairness to the ISOs (and as shown in Chapter 4), the design, creation, and operation of real-time markets is a complicated undertaking.

EXAMPLES OF ISO PROBLEMS

The ISOs make day-ahead (as well as hour-ahead) load forecasts to help decide how much generating capacity to acquire for each operating hour. These forecasts will likely differ from the day-ahead (and hour-ahead) schedules submitted by suppliers and load-serving entities and will likely differ from the loads that occur in real time.

If the ISO load forecast exceeds the scheduled demand, should the ISO acquire additional resources? If it accepts the market schedules and higher loads materialize in real time, reliability may be threatened. On the other hand, if the ISO acquires additional generating capacity to meet the difference between its load forecast and the schedules and the ISO's forecast is too high, MCPs will be lower, and electricity costs will be higher than if the market was permitted to operate without ISO intervention; see the example on low-operating limit in Chapter 4.

Between June and October 2000, the California ISO's day-ahead forecast exceeded actual hourly loads by an average of 300 MW (Fig. 18). Specifically, the ISO's forecast was

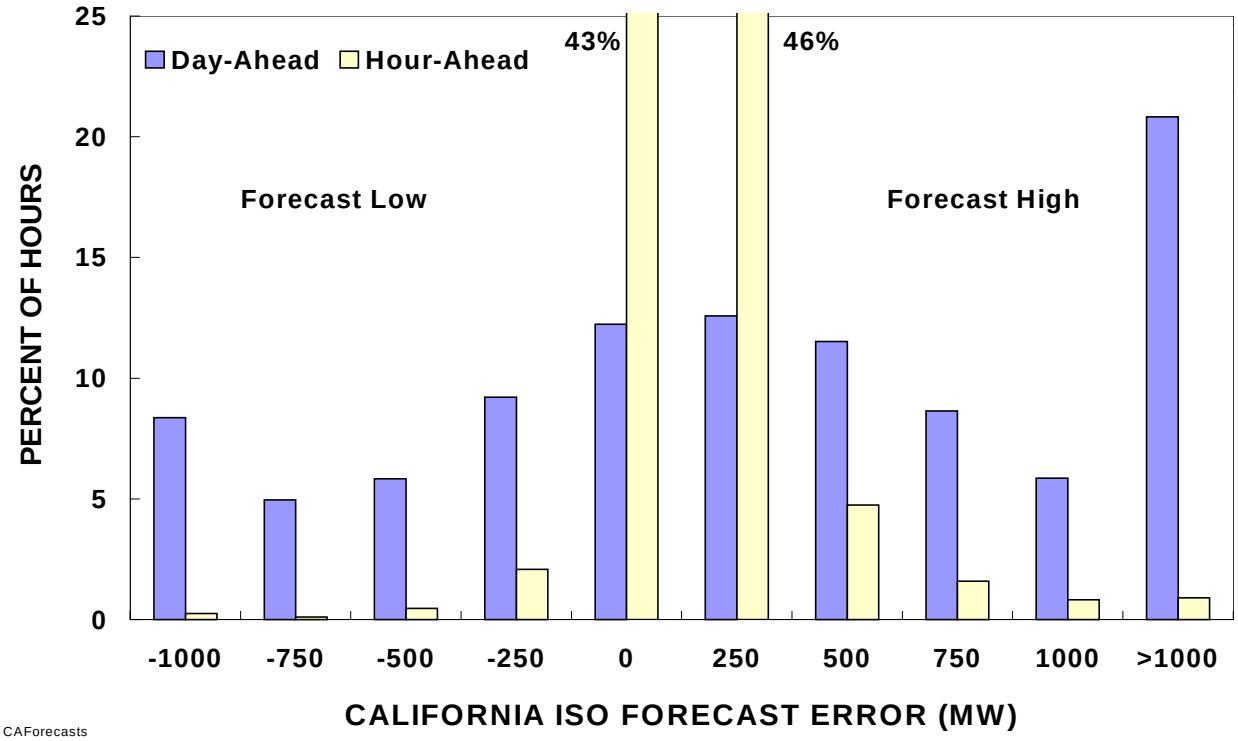


Fig. 18. Accuracy of day- and hour-ahead load forecasts by the California ISO.

higher than the hourly load by 1000 MW or more for over 20% of the hours. (The ISO's hour-ahead forecasts were, as shown in Fig. 18, much more accurate than the day-ahead forecasts.) On the other hand, the day-ahead schedules were, on average, about 2500 MW below the actual loads. These differences illustrate the dilemma an ISO faces in trying to decide whether to favor reliability or markets. When the ISO forecast exceeds the market schedules, the ISO typically purchases additional replacement reserves to ensure the availability of sufficient capacity in real time to meet loads.*

For a variety of reasons, including overforecasting of actual loads, ISOs often acquire resources "out of market." Such acquisitions do not set the MCP; instead, their costs are collected through an uplift charge. Chapter 4 presents examples in which such charges occur.

ISO New England has experienced growing problems with such energy-related uplift charges. In 1999, the ISO spent $73,000 a day on uplift, primarily related to units operating at the LOL; in 2000, the comparable expense was almost $320,000, more than four times higher. Figure 19 shows how these uplift charges varied from month to month. Apparently, the New England market rules implicitly encourage generators to bid unit characteristics that are much less flexible than actual unit characteristics. This pay-as-bid feature has depressed the MCP for

*Between June and November 2000, ISO New England overforecast hourly loads by an average of 110 MW (about 1% of load). The ISO overforecast hourly loads by 500 MW or more for 8% of the hours and underforecast loads by 500 MW or more for only 1% of the hours.

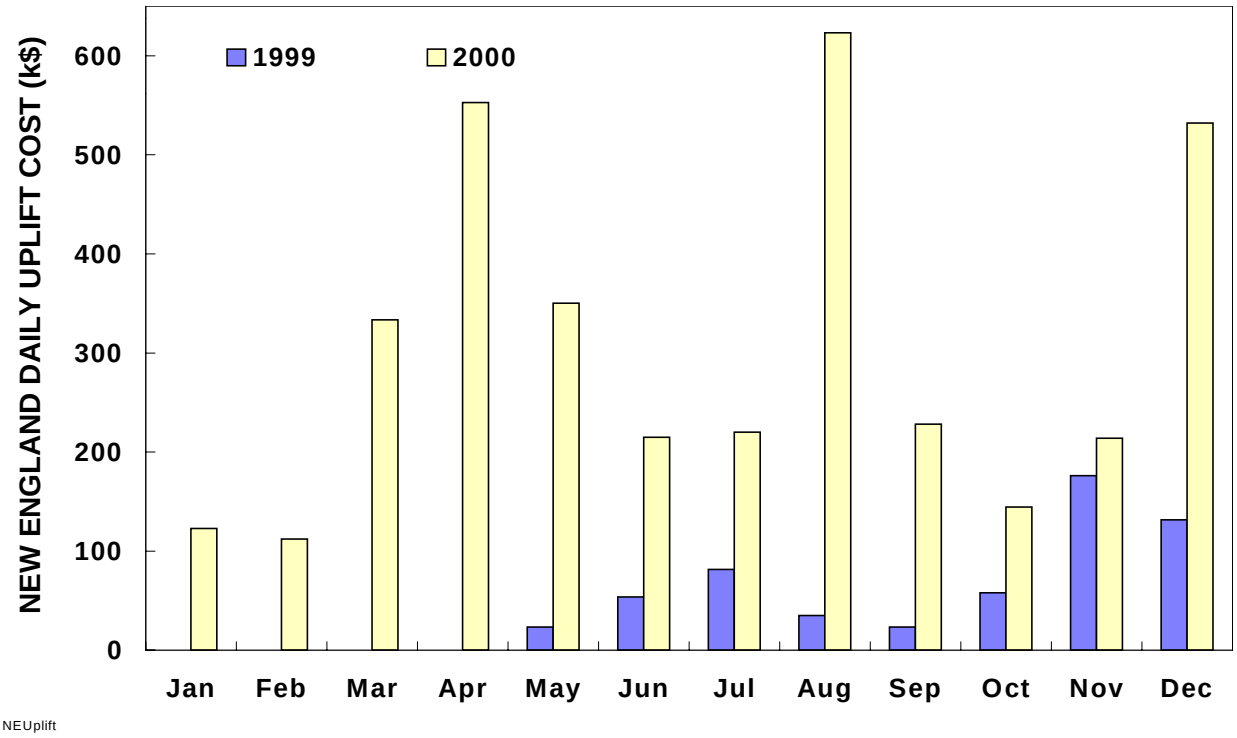


Fig. 19. Energy-uplift charges paid by ISO New England from May 1999 through December 2000.

energy and reserves (by forcing inflexible generators online and requiring less expensive units to back down), which has discouraged construction of flexible generating units that can respond quickly to price spikes.

RTO PLANS

The RTO filings of October 2000, required by FERC's Order 2000, pay scant attention to real-time markets and operations. Perhaps because of the need to resolve other RTO issues, such as governance, regional scope and membership, and transmission cost allocation and revenue requirements, the filings largely ignore the important and complicated issues associated with intrahour markets. (Some of the RTOs plan to retain multiple control areas, in which case the intrahour operations and markets discussed here are not relevant. In such cases, the Order 2000 balancing requirement defaults to an hourly accounting function equivalent to the energy-imbalance ancillary service specified in FERC's Order 888.)

For example, the RTO West plan (Avista Corp. et al. 2000) includes an "Attachment N: Description of RTO West Ancillary Services." One of the eight ancillary services RTO West plans to offer is Balancing Energy, defined as:

> RTO West's coordinated use of Regulation, Load Following Up, Load Following Down, Replacement Reserve and Supplemental Energy resources and,

> to a limited extent, Spinning Reserve and Non-Spinning Reserve resources (for the period of time during which these two types of resources are dispatched in response to a contingency) — in real-time to deliver energy to, or acquire energy from, each SC's account in order to balance each SC's account on a periodic (ten-minute) basis and to enable the RTO West to comply with NERC and WSCC [Western Systems Coordinating Council] control area performance standards.

This is the sum total of RTO West's plan for an intrahour balancing market.

Although the Midwest ISO (2001) claims that it "readily satisfies all four of the minimum characteristics and eight functions of Order No. 2000," it has no concrete plans to create and operate a real-time balancing market. Its filing states that:

> The Midwest ISO and the Transmission Owners are currently working in tandem with other interested market participants to develop a Midwest ISO Schedule 4 — Energy Imbalance Service. … Sorting out the options for determining the market clearing price within an hour has proven to be somewhat problematic for the Midwest ISO. As the Midwest ISO did not evolve from a tight power pool and currently does not operate a pool, it has been a challenge for the Midwest ISO to develop a mechanism for determining the market clearing price.

To the extent the Midwest ISO considered FERC's balancing-market requirement, it dealt with settlements rather than operations and with hourly pricing rather than intrahour markets.

Finally, the GridSouth filing (Carolina Power & Light et al. 2000) proposes a traditional Energy Imbalance Service, complete with deadbands and penalties. (This service is consistent with FERC's Order 888, which treats energy imbalance as a cost-of-service function rather than a competitive function, as outlined in Order 2000.) Its filing, while recognizing that "the proposal does not establish a real-time balancing market," is in my view a step backward because its imbalance proposal ignores completely market forces in real-time operations and pricing. GridSouth offers three reasons for its failure to develop a real-time balancing market as of the Independence Date: (1) creation of such a market to span three control areas would require substantial investment; (2) development of such a market must be coordinated with the development of other markets for congestion, energy, and ancillary services, and such development should be managed by the new GridSouth entity; and (3) GridSouth should build on experience in other regions where real-time pricing issues are being addressed.