Review of the Proposed Reserve Markets in New England*

Peter Cramton, Hung-po Chao, and Robert Wilson
University of Maryland, EPRI, and Stanford University
18 January 2005

Contents

* We have benefited from the helpful comments of Gail Adams, Mario DePillis, David LaPlante, and Jim Milligan. The views expressed are our own.

Review of the Proposed Reserve Markets in New England

Peter Cramton, Hung-po Chao, and Robert Wilson
18 January 2005

1 Summary

ISO New England proposes reserve markets designed to improve the existing forward reserve market and improve pricing during real-time reserve shortages. We support all of the main elements of the proposal. For example, we agree that little is gained by allowing reserve availability bids in the day-ahead market. Doing so greatly increases the complexity of the market without the prospect of more efficient pricing. Rather, offline reserves are most efficiently priced and awarded well in advance, as is done by the improved forward reserve market.

1.1 Background

The goal of the reserve markets is to motivate efficient investment and operation of resources to provide energy on an emergency basis. This objective has proved challenging in all restructured electricity markets. The challenge arises from non-convex costs (large fixed costs and negligible variable costs), lumpy investment and commitment decisions, asymmetric information, and local market power. The challenge has been especially great in New England because of large first and second contingencies, and hence a greater need for reserves to maintain system security.

In March 2003, New England began operating under its standard market design (SMD), a multi-settlement energy market with locational energy prices. Reserve markets initially were ignored in order to implement SMD as quickly as possible. However, it was envisioned that reserve markets would be introduced later. The initial plan was to extend the day-ahead and real-time energy markets to include reserve availability bids. The spot markets then would jointly optimize the schedule and dispatch to satisfy energy and reserve constraints. Both energy and reserves would be priced and scheduled day ahead, with the real-time market used to settle deviations.

This approach was ultimately rejected. Not only is such a market complex for participants and difficult for the ISO to implement, it would not motivate efficient investment and operation of reserve resources. The primary difficulty is that in both the day-ahead and real-time markets,

the marginal cost of providing reserves is negligible. Hence, absent market power, reserve availability bids will tend to zero. The result is negligible prices for reserves and an inadequate incentive to invest in valuable reserve capacity.

To see the problem, consider an import-constrained zone with too few fast-start resources for offline reserves. As a result, the ISO must commit additional non-economic resources to provide sufficient reserves. These resources are then paid (via uplift[1]) the portions of their commitment costs that are not recovered from energy and reserve sales. The extra quantity of energy and reserves provided by these non-economic unit commitments can depress both the energy and reserve prices. The result is less incentive to invest in fast-start resources, just the opposite from the desired response. In the short run, this dearth of fast-start resources is likely to reduce system flexibility and reliability. In the long run, this will induce the technology mix to shift away from peakers towards baseload units, increasing the long-run cost to consumers.

1.2 Proposed reserve markets

The forward reserve market and shortage pricing in real time form the core of the proposed reserve markets. Instead of extending the day-ahead and real-time markets to include reserves, the ISO proposes shortage pricing in real time, in addition to enhancing the forward reserve market for offline reserves.[2] These instruments address more directly and simply the three basic market problems: (1) system-wide shortages of operable capacity, (2) local shortages of operable capacity, and (3) an inefficient mix of offline and online reserve resources.

1.2.1 Real-time reserves

To address the problem of inadequate operable resources, real-time shortage prices are determined from a penalty factor for each type of reserve. During reserve shortages, these prices are paid to all resources providing energy or reserves (except those resources with a forward reserve obligation)—a bonus for being available when they are most needed. When there is a shortage, high shortage prices will motivate increases in supply and decreases in demand.

[1] The current market rules call this cost "Operating Reserve Credits." It is charged to load on a pro-rata basis.
[2] We use the term "shortage pricing," rather than "scarcity pricing," since all prices are about scarcity; whereas, we are referring to pricing during a shortage of reserves.

1.2.2 Forward reserves

The forward reserve market directly addresses the mix of reserve resources. All resources, both offline and online, compete in an annual (or semiannual) auction to supply offline reserves. Auctioning a long-term product has many advantages: (1) entry and exit are possible, enhancing competition, (2) fixed costs become variable, and thus are more accurately represented in the price, (3) prices are less volatile, and (4) planning is facilitated.

The definition of the forward reserve product is critical. Forward reserve resources are obligated to offer energy in the day-ahead and real-time markets at or above a floor price, which is set so the resources are rarely called for energy. There are penalties for failure to be available and failure to perform when called. Forward reserve resources do not receive the shortage price, nor do they receive operating reserve credits to cover commitment costs. This definition applies whether the reserves are provided by offline or online resources. In this way, online and offline resources can compete on an equal basis to provide offline reserves. Online resources will include the forfeited commitment costs in their bids. Performance penalties encourage units to be more reliable. In this way, the forward reserve market encourages least-cost supply of offline reserves by self-selecting resources with lower energy opportunity costs, higher reliability, and lower commitment costs.

To further facilitate the efficient substitution of offline and online resources, the forward reserve market allows portfolio offers (from those with multiple resources) and virtual offers, which become physical in the day-ahead market. The ISO also intends to accommodate the bilateral trade of forward reserve obligations. Participants short on their obligations can buy additional supply on a daily or long-term basis. However, the forward reserve market is not settled against a real-time reserve market. This would not make sense, since the forward reserve product is different from the real-time reserve product. For example, a unit providing forward reserve forfeits compensation for commitment costs, is ineligible for shortage bonuses, and faces penalties for failing to be available or perform when called.

1.2.3 Relationship to locational capacity market

ISO New England has proposed a locational capacity market (LICAP) to provide the right investment incentives in New England. Since both will be present, LICAP and the forward reserve market must work together as complements. In the absence of LICAP, the forward

reserve market has proved essential in motivating the supply of flexible resources. Following the introduction of LICAP, the forward reserve market will play a less important role in providing incentives for flexible resources, since these resources will be substantially rewarded by LICAP. However, the forward reserve market will continue to provide additional compensation to reserve resources to the extent they are undersupplied in a particular location. Most importantly, the forward reserve market sets a locational price for reserves well in advance of the spot market. In this way, the price reflects the economic costs of reserve supply from units other than quick start generators, such as dispatchable load or slow start units that provide reserves from ramping. Thus, although we can expect forward reserve prices to drop substantially with the introduction of LICAP, the forward reserve market will provide supplemental compensation when and where flexible resources are scarce.

1.2.4 Recommendations

Shortage pricing

The ISO's shortage pricing proposal recognizes the high value of energy and reserves in times of shortage. The high spot price during shortages motivates demand to reduce consumption and encourages suppliers to capture the high spot price through greater availability. We make the following recommendation with respect to shortage pricing.

Shortage prices should be paid to all energy and reserves. Energy resources should receive the shortage price in the real-time energy price. The day-ahead energy price then would reflect the expected shortage price and provide an efficient basis for day-ahead commitment and scheduling. Suppliers, absent market power, would still have an incentive to offer their resources in the day-ahead market at the marginal energy cost. To the extent that the day-ahead price does not fully reflect the expected shortage price, suppliers could submit virtual bids to shift their energy sale to the real-time market, while preserving efficient day-ahead commitment and scheduling.

Shortage pricing may introduce incentives to create real-time shortages. This is a highly destructive exercise of market power, since creating real-time shortages undermines reliability. This potential problem is avoided by deducting the shortage price from all LICAP resources, as is currently proposed. Then resources still face the same marginal incentive to provide energy or reserves during shortages. All LICAP resources have the shortage price deducted from revenues,

but those that are available receive the energy and reserve prices to offset this deduction. There is no incentive to create shortages, since the net gain from creating a shortage is zero. The increase in energy and reserve compensation is completely offset by the deduction in LICAP revenues. LICAP then provides a hedge of the shortage price for load and improves bidding incentives in the day-ahead and real-time markets.

The need for a performance-based LICAP product is clearly illustrated by the events of 15 January 2004. Although ICAP resources were plentiful, the system nearly collapsed because many of the ICAP resources were unavailable due to a lack of gas. With the performance-based LICAP product, as currently proposed, resources are only able to capture LICAP and peak-price revenues by demonstrating availability at these critical times. This is an essential innovation to the capacity market.

Reserve availability

With the introduction of shortage pricing, it is important to make sure that reserve availability is accurately measured. Resources should only receive shortage prices for providing reserves based on a demonstrated track record of availability. This prevents unreliable resources from getting shortage prices by submitting extremely high energy bids, so that they are never called. The forward reserve market includes an auditing procedure for demonstrating availability. This same approach should be used for resources that are providing real-time offline reserves. The proposed LICAP market also intends to use the same auditing procedures for the measurement of offline reserve availability. In this way any offline resource capable of providing reserves is evaluated in the same way whether they choose to be rewarded for their reserve capability in the forward reserve market, the real-time reserve market, or the LICAP market.

Demand-side participation

The absence of demand response has been a major shortcoming in all electricity markets. This absence will never be corrected if demand-side participation is limited. Hence, it is important that dispatchable load be able to participate in the reserve market. Dispatchable load to the extent that it provides reserves also should be eligible to be a LICAP resource, further enhancing the economic incentive for demand-side participation.

Next steps

Important next steps in the design of these markets are to determine: (1) the performance penalties in the forward reserve market, and (2) the penalty factors for shortage pricing in real time. Market simulation will be useful in examining the likely consequences of the alternatives.

2 Introduction

We were asked by ISO New England to review the design of the proposed reserve markets in New England. The proposal has been under development throughout the period of our study, which began in late 2003. Numerous alternatives have been considered and rejected. Here we discuss the alternatives as well as the final proposal as outlined in the ISO New England white paper, "Ancillary Service Markets Enhancements," 6 May 2004, and subsequent documents on the proposed LICAP market. Our objective is to critique the proposal as well as identify key issues for further consideration and testing.

Reserve markets remain one of the major challenges of electricity market design. The basic difficulty is how to motivate efficient investment in resources that stand ready to provide emergency energy but are rarely called to generate energy. Non-convex costs are part of the problem—lumpy investments with high fixed costs and low variable costs—but performance incentives are also important in light of asymmetric information and the fact that reserves are only rarely called to provide energy.

Our analysis assumes that the objective of the design is an efficient and reliable energy market. Energy and reserves should be supplied to satisfy demand and all reliability constraints at least cost. Moreover, the market should send price signals to encourage the right mix and magnitudes of efficient (least cost) investments in new capacity.

We begin with a description of the treatment of reserves in the current markets, followed by a general analysis of reserve issues. We then analyze the key elements of the proposed design. We conclude with our recommendations. Appendices present a mathematical formulation of the proposed design and more detailed discussion of certain issues.

3 Description of the markets

3.1 Reserve requirements

Current reserve requirements are as follows:

- Ten-minute spinning reserve (TMSR): ½ of first contingency.
- Ten-minute non-spinning reserve (TMNSR): ½ of first contingency.
- Thirty-minute operating reserve (TMOR): ½ of second contingency.
- Thirty-minute replacement reserve: ¼ of second contingency.

Day-ahead reserve requirements are based on the day-ahead forecast of first and second contingencies. Day-ahead locational requirements also are based on forecasts. The reliability adequacy assessment requirements are based on the ISO's forecasts of real-time requirements. Real-time requirements are based on actual conditions, and real-time locational requirements are based on a security analysis of real-time conditions.

Reserve requirements for the New England system are stable across the year (and especially within each season), since the first and second contingencies rarely change. The largest contingencies come from the DC line from Quebec (between 1200 and 1500 MW) and two similarly sized nuclear units (each about 1200 MW). Local reserve requirements vary from hour-to-hour based on local conditions.

All physical resources that bid into the day-ahead market are considered available for reserves. Availability is calculated from the response rate and unused capacity for spinning resources and claim10 and claim30 for non-spinning resources (the MW amount of capacity available from an offline resource in ten and thirty minutes, respectively). Dispatchable loads are eligible for non-spin (and spin if and when there is a change in the NERC criteria).

3.2 The existing day-ahead and real-time markets in New England

This section explains the mechanics of bidding into the current day-ahead and real-time markets. The day-ahead market is optimized to meet energy and reserve requirements at least cost. Inputs include supply offers, demand bids, virtual bids (INCs and DECs), but no reserve availability offers. The day-ahead market creates financial obligations for energy, but not reserves. The real-time market has the same objective function as the day-ahead market, but the real-time market is translated into a physical dispatch. Virtual offers and bids (INCs and DECs)

are not allowed. The clearing is based on actual system conditions. The real-time energy market then settles based on deviations from the day-ahead market.

ICAP units have an obligation to bid in the day-ahead and real-time markets. Bids are submitted by 1200 (noon) day ahead. Each unit offers a single energy schedule with up to ten blocks (either steps or piecewise linear). The same offer schedule applies to every hour of the day. Reserve constraints are included in the day-ahead optimization, but there are no reserve offers, either day ahead or in real time. Physical characteristics are also bid: ramp rate(s), startup and no-load costs, lower and upper operating limits (EcoMin and EcoMax), and for offline resources, claim10 and claim30. The day-ahead energy market includes virtual offers and bids (INCs and DECs) that enable participants to arbitrage expected energy price differences between the day-ahead and real-time markets. Shortly after 1200, the day-ahead market is cleared and the resulting day-ahead schedule is posted. Currently about 90 percent of the load is bid day-ahead. About 60 percent of the generation is self-scheduled; a substantial fraction of the self-scheduled generation is from units with start times greater than 8 hours, such as nuclear units. The local second contingency is ignored. The day-ahead schedule is determined by a dispatch security-constrained to satisfy the demand as bid, plus reserves sufficient to cover the operating reserve requirements for the New England region.

At 1600, a two-hour re-offer period begins. This allows resources that did not clear in the day-ahead market to bid; no demand bids or virtual bids are accepted—only physical supply can bid in the real-time market. Resources that cleared day-ahead cannot change their offers in the real-time market. At 1800, the reliability adequacy assessment (RAA) is run to translate the bids into a physical real-time dispatch, including scheduling of supplemental reserve capacity. The RAA is rerun about every four hours during the day.

The objective of the RAA is different from that of the day-ahead and real-time markets. The RAA minimizes the cost of getting sufficient resources online to satisfy the load forecast plus reserve requirements. Forecasted load that was not bid in day ahead is treated as additional online capacity requirements. However, because about 90 percent of load is bid in day ahead, the typical action as a result of the RAA is to schedule additional RMR units to satisfy the local second contingency in constrained zones. This step tends to commit more expensive resources,

since it does not consider the energy price for units that are committed for reserves, but will ultimately provide energy.

3.3 Proposed forward reserve market

The ISO awards all offline reserves (TMNSR and TMOR, expanded to include replacement reserves) in an annual (or perhaps semiannual) uniform-price auction. Local reserve constraints are included for import-constrained zones, based on peak conditions. Market clearing allows superior products (TMNSR) to substitute for inferior products (TMOR); hence, the price of TMNSR is at least as great as TMOR, and the price in a constrained zone is at least as great as the price in an unconstrained zone. Portfolio bids and virtual bids (within a zone) are allowed, but the resources that can physically provide the reserves must be declared before or during the day-ahead market. The auction determines each bidder's award or delivery requirement for the various products and the clearing price that is paid in each hour of delivery. The delivery requirement can be traded in an ISO-facilitated bilateral market before the day-ahead market.

The delivery requirement is satisfied by: (1) designating a physical resource to provide reserves in the day-ahead market; (2) offering the required number of MWs in the day-ahead and real-time energy markets at a price at or above the specified floor price (the offered MWs must be available as energy in ten or thirty minutes, depending on the product); and (3) providing the required MWs as energy when called. Offline reserves can be provided by fast-start offline resources or online resources (as spin). Those satisfying the delivery requirement are paid the forward reserve (FR) clearing price (and the energy clearing price if called for energy). FR resources do not receive the real-time capacity adder (shortage price). Online FR resources are ineligible for commitment cost compensation.

Failure to satisfy the delivery requirements results in forfeiture of the FR payment. In addition, a penalty is assessed, which depends on the type of failure and the state of the system. There are three types of failures (in increasing order of severity): (1) failure to be available declared during the day-ahead market, (2) failure to be available declared after the day-ahead market, and (3) failure to perform when called for energy. The formulas for the penalties are still under review.

3.4 Shortage pricing

In the event of a shortage of reserves (either system-wide or locally), penalty factors determine a shortage price that is paid to all non-FR resources that are providing energy or reserves. The penalty factor depends on the type of shortage. The shortage price is higher for shortages of higher quality reserves, and higher quality reserves can substitute for lower quality. Thus, quantities cascade down to lower qualities and prices cascade up to higher qualities. Locational reserve constraints are included. Hence, the local shortage price may be positive, indicating the violation of a local constraint, even when there are plenty of reserves system-wide.

The penalty factors are chosen to be consistent with the actions system operators would take to restore real-time reserves if such offers were available. Since under current NPCC rules, ISO-NE may operate short of TMOR for up to four hours, the price for a shortage of less than four hours is modest, but for shortages greater than four hours, the TMOR shortage price is high, approaching the energy price cap of $1000. In contrast, system operators are required to maintain ten-minute reserves to cover the first contingency. Hence, the shortage price for TMSR and TMNSR is close to the energy price cap of $1000 in all circumstances.

4 General reserve issues

Reserves are an essential product in reliable electricity systems. Reserves enable the system operator to promptly correct a supply-demand imbalance. Without reserves, the loss of one or more large generators or transmission lines could lead to a catastrophic failure of the system. Despite this obvious need, reserve markets have proven extremely difficult to implement. Early attempts suffered from both poor product definition and poor market design. Here we discuss the basic issues in designing a reserve market.

4.1 Reserves are a public good in current electricity markets

Unlike energy, reserves are a public good on the demand side.[3] Reserves improve reliability for all load. For this reason, reserves should be centrally procured by the ISO. Load should not participate directly as a buyer of reserves. If reserves were procured on a voluntary basis like energy, then any individual load would free-ride on the reserves of others. Hence, required reserves cannot be procured efficiently on a voluntary basis. Rather, load's direct participation in the reserve market should be on the supply side, offering reserves as dispatchable load.

In contrast, energy is a private good, and thus the justification for direct demand-side participation in the energy market is strong. Load is constantly making decisions about how much energy to consume. Efficiency requires that load be able to express its willingness to pay for energy.

4.2 Reserves as emergency energy with performance incentives

Most prior markets have defined reserves as idle capacity with the claimed capability of producing a specified quantity of energy within a particular time frame (ten or thirty minutes). A better definition for reserves is emergency energy—a physical option on energy callable in a specified time frame (ten or thirty minutes) in response to a contingency.

The main difference between the idle capacity and emergency energy approaches is what happens in the event of nonperformance. With reserves treated as idle capacity, a failure to perform results in a loss of the reserve payment and the lost opportunity of providing energy at a

[3] This statement rests on the incompleteness and imperfections of current electricity markets. One can theoretically imagine a world where reserves are a private good for both sides of the market, but our markets have not evolved to this state.

high price. In contrast, with reserves treated as emergency energy, there are penalties for failing to be available and a larger penalty for failure to perform when called for energy.

Performance penalties are necessary to provide efficient performance incentives. Without a penalty, the generator owner would have no incentive to disclose to the ISO that a particular unit is inoperable. The owner would prefer to collect the reserve payment. Instead, penalties should be structured so that the owner does have the incentive to report the inoperable condition of the unit as soon as the condition is known. Then the ISO can better assure that it has sufficient operable reserves. In addition, performance penalties give the owner proper incentives to make investments to improve the reliability of its units. These investments include regular testing and maintenance, configuration to handle multiple fuel types, and the purchase of firm gas in winter months.

The penalty should depend on the size of the harm caused by nonperformance. For example, declaring a unit inoperable before the day-ahead market means that the ISO has time to make alternative arrangements in the day-ahead schedule. Finding alternative resources becomes increasingly difficult as we approach real time. Thus, the penalty for a failure to reserve capacity should increase as we get closer to real time. A second and more costly form of nonperformance is the failure to provide energy when called. The harm to the system depends on the scarcity of resources.

With the emergency energy approach, bidders must put the expected cost of nonperformance in their bids. More reliable units can afford to underbid less reliable units. As a result, the emergency energy approach encourages market participants to offer more reliable and better maintained resources as reserves. In contrast, under the current definition, all idle capacity that can produce energy in ten or thirty minutes *is* considered to be providing reserves. Even though the quality of reserve varies from unit to unit, both reliable and unreliable offline units receive the same compensation.

Reserves should not be defined as a standard call option on energy at a specified strike price. First, for reliability, it is essential that reserves be physical, not purely financial. Second, reserves are a reliability instrument, not a hedging instrument for load (unlike a pure call option on

energy).[4] For this reason, it makes sense for the reserve supplier to be paid the energy price in the event the unit is called, rather than a strike price. Receiving the higher energy payment induces the supplier to offer reserves at a lower price.

4.3 Long-run resource mix

A fundamental goal of reserve markets is to send the right price signals to motivate investment in the efficient mix of resources. Reserves can either come from fast-start (offline) resources or from online resources with unloaded capacity. An efficient reserve market will generate prices that provide the right balance between online and offline resources; that is, offline resources will be favored if and only if they provide services at lower cost than online resources.

A major challenge of the reserve market is to produce prices that reflect the cost of service. The problem is that many of the fixed costs of reserves are not adequately recovered by prices in the spot markets. This issue is discussed in the following section.

4.4 Spot vs. forward purchase of reserves

There are two basic approaches to reserves. One emphasizes the spot market; the other emphasizes forward purchase.

The goal of the spot approach is to determine the right spot price for reserves in a simultaneous optimization with energy. This spot price is intended to promote efficient long-run investment and efficient short-run dispatch of resources. A forward market may also be included to hedge spot price volatility, but the system relies primarily on the spot market for efficient pricing and investment incentives.

The forward approach involves purchasing reserves well in advance on a long-term basis. A spot market may be used to rebalance forward positions, but the forward prices drive investment decisions.

4.4.1 The spot approach to reserve markets

We first consider the spot approach in greater detail. The idea behind this approach is that reserves are not much different from energy. However, energy and reserves have important

[4] Energy options may be desirable products for load to purchase, but the purchase of such options should be done by load, perhaps as mandated by the appropriate Public Utility Commission, and not by the ISO. See Chao and Wilson (2004), "Resource Adequacy and Market Power Mitigation via Option Contracts", EPRI, Palo Alto, CA.

economic differences that make it difficult for a spot market for reserves to allocate resources efficiently (see Table 1).

Table 1. Differences between energy, online reserves and offline reserves

Issue	Energy	Online Reserves	Offline Reserves
Public or private good	Private good for consumption	Public good for reliability	
Demand quantity	Demand varies each hour	Demand is largely known in advance	
Marginal cost	MC is substantial and varies in both DA and in RT	MC is zero in RT, but may include unit commitment costs DA	MC is zero both DA and in RT
Selection of optimal provider	Large variation in optimal provider both in DA commitment and RT dispatch	Optimal provider largely determined by DA commitment	Optimal provider only rarely determined by DA commitment and RT dispatch

A main difference between energy and reserves is that energy marginal costs are significant and are different for each resource. This means that the real-time dispatch is critical to short-run energy market efficiency. Resources also have startup costs, no-load costs, and other constraints, which make the day-ahead commitment of resources important for short-run energy market efficiency. Finally, because energy marginal costs are significant and varied, there is scope for inframarginal units to earn substantial rents, which can contribute to fixed costs.

In contrast, reserves in most situations have zero marginal costs in real time. The exception is when online units are backed down to provide reserves, but in this case the energy opportunity cost (energy price minus energy bid) is easily calculated and compensated.[5] Assuming marginal cost bidding, the reserve price then will be zero in real time, except in rare instances when there is a reserve shortage.

Even if day-ahead availability bids were allowed, assuming that it is possible for participants to arbitrage expected price differences between the day-ahead and real-time markets, the day-ahead price would be the expected real-time shortage price. If the day-ahead price were to

[5] Offline units may occasionally be held back to provide reserves as well. The typical case is a limited energy hydro unit. Reserves are maximized by bringing a fossil unit on line to satisfy energy, leaving the hydro unit to provide reserves. Because of the hydro unit's inability to sustain energy output over an extended period, it is a better source of reserves.

exceed the expected real-time price, then it would be profitable for a participant to sell additional reserves day ahead, and then fill the additional reserve obligation with real time purchases.

Unlike the day-ahead energy market, a day-ahead reserve market would not play an important role in the efficient scheduling and dispatching of resources. The efficient scheduling and dispatching of reserve resources follows from the three-part energy bids and the physical characteristics of the units. Because there are negligible day-ahead costs, a day-ahead reserve market would be simply an opportunity for participants to make bets on the likelihood of a shortage in the next day. Absent expected shortages, the day-ahead reserve offers would be zero, both from online thermal resources and offline fast-start units. Hence, the justification for bid-based day-ahead and real-time reserve markets is weak.

4.4.2 The forward approach

The alternative to the spot market approach is for the ISO to purchase offline reserves on a forward basis, as is done currently, and which the ISO proposes to continue. Forward purchase has numerous advantages relative to the spot approach: 1) Reserves can be acquired at a point in time when most costs are variable rather than fixed. 2) Competition is enhanced, since entry and exit become possible. 3) Market-based prices are determined without resorting to administratively set penalty factors, which may distort investment decisions in both the reserve and energy markets. 4) Finally, forward purchase is much simpler both for the ISO and for participants. Participants do not need to estimate spot prices, nor decide their bidding strategy on a daily basis. Suppliers learn of their obligation well in advance and then can determine the most efficient means to satisfy the obligation, such as purchase, staffing, and maintenance decisions.

The approach involves the ISO specifying a quantity of reserves to be purchased in advance, and the duration of the contract. In New England, the reserve requirement does not vary much over the season or the year, since it depends on first and second contingencies, which rarely change. Hence, there is little uncertainty in the quantity of reserves that needs to be purchased. The purchase of reserves via auction can occur every six months (as is done now), every year, or for multiple years. Recent experience with long-term forward markets has been promising;

examples include the basic generation service auctions in New Jersey and the capacity auctions in France and Belgium (all involve contracts with a duration of up to three years).[6]

The current forward reserve market in New England is physical and does not allow bilateral trading. These restrictions were made to expedite implementation and can be relaxed, as ISO-NE proposes. Offers, however, must specify pricing location, in order to address locational constraints. In addition, providers must declare physical resources available at the time of the day-ahead market to assure that reserve constraints are satisfied. By allowing and processing bilateral trades, the ISO can improve market efficiency. A lower-cost supplier is able to replace a higher-cost supplier. For example, a hydro resource that normally provides reserves, because of its limited energy capability, may find that it has excess water, and can more efficiently provide energy. This would be true if there is another resource that can profitably substitute for the hydro unit. Absent market power, all voluntary bilateral trades improve efficiency.

The forward purchase of reserves is consistent with the practices of other low marginal cost, non-storable commodities. For example, internet access typically is priced on an annual basis (often with monthly payments), as is cable service, wire line phone service, and most wireless phone service. Spot pricing based on network congestion is not done. For other capacity-limited commodities, such as hotel rooms, rental cars, and airline seats, prices do respond to supply shortages, but primarily as a means to manage demand. This is not an issue for reserves, where demand is given by a fixed requirement. Load cannot reduce demand for reserves in response to high prices.

Locational forward reserves

It is non-trivial to introduce locational constraints in a forward reserve market. The reason is that the locational constraints depend on the peak flows, which depend on the system conditions. Although system-wide reserve constraints are roughly constant, locational reserve constraints are much more variable.

The simplest approach is to add locational constraints based on peak local reserve requirements. This may result in too much local reserve in off-peak periods, but reserves for the

[6] Lawrence M. Ausubel and Peter Cramton, "Auctioning Many Divisible Goods," *Journal of the European Economic Association*, 2, 480-493, April-May 2004.

overall system will be correct. This solution may favor offline technology in load pockets, but this is probably a good thing given the pressing need for offline resources in load pockets.

4.5 Motive for a forward reserve market

Reserves have an especially important role in New England. ISO-NE has the usual need for reserves for load following (to meet load surges) and to ease routine reliability problems (from occasional transmission and generation interruptions). In addition, the New England system needs substantial reserves for system security. These reserves are necessary because it is vulnerable to large first and second contingencies associated with the DC line from Quebec and three large nuclear generators. In comparison with other regions, the need for quick-start units is great and the supply is limited, as shown in Table 2. Currently, ISO-NE must often rely on online resources to meet offline reserve requirements, especially in import-constrained zones. This is not surprising, given that compensation for offline reserve units historically has been inadequate to cover costs. The problem with this outcome is that New England's resource mix likely is inefficient.

Table 2. Quick-start units by region*

	ISO-NE	NYISO	PJM
Peak load	25,000	31,000	54,200
First contingency	2,000	1,200	1,157
Second contingency	1,160	1,140	1,157
Total requirement	2,580	1,770	1,736
Quick-start units	1,500	3,000	8,000
Quick-start/requirements	58%	169%	461%

*Offline ten-minute resource capacity and total reserve requirements as of 2001.

Our focus here is on offline reserves, including 10-minute non-spin, 30-minute operating reserves, and replacement reserves. One-half of the first contingency is met with non-spin; also, one-half of the second contingency is met with 30-minute operating reserve and one-quarter with 30-minute replacement reserve. Online resources can compete in the spot markets for the provision of these reserves, even though they are designated as "offline." More generally, spin can substitute for all lower qualities of reserves.

ISO-NE recently introduced a semiannual forward market for offline reserve requirements. The chief motive was to improve the resource mix by paying for valuable services. Fast-response offline resources (mainly internal combustion units (ICUs)) would be more likely to obtain

sufficient revenue to cover their capital and maintenance costs, and the ISO would reduce the uplift from unrecovered costs of additional unit commitments. In the following section, we explain our understanding of the origin of the problem, and the role of the forward market in addressing this problem.

4.5.1 Origin of the problem

One could argue that in principle the spot markets (day-ahead (DA) and real-time (RT)) might provide sufficient revenues to attract the right amount of offline resources. This has not been the case in New England because of the pricing rules in its spot markets. An offline resource's reservation price for capacity availability, interpreted as opportunity cost, is forced to zero in the RT market; thus, revenues are essentially the product of the frequency of energy dispatch during reserve shortages, the average dispatch duration, and the energy price.

The value of offline resources is also not fully represented in the DA market. Offline resources are substitutes for online units (with surplus capacity, or backed down to release capacity).More importantly, offline resources substitute for the commitment of additional thermal units to ensure sufficient capacity to meet the excess of scheduled load over forecasted load. This includes additional units committed to provide spin available to meet under-scheduled load, and thus typically with surplus capacity available for operating reserve.

The key feature of the settlement rule is that the cost (startup and other costs not recovered from energy sales) of unit commitments is uplifted, and therefore not reflected in the scarcity value of reserves. Moreover, the additional commitments depress the energy price, further reducing potential revenues of an ICU (either from energy or energy opportunity cost). Because the full value of offline resources is not reflected in settlements, fast-response peaking units such as ICUs often obtain insufficient revenues to cover their full costs. Indeed, the deficiencies of the settlement rule could lead to a vicious cycle. Under-pricing of offline resources leads ICUs to leave the market, which requires the commitment of more resources to meet reserve requirements, which further depresses the price for reserves from offline units because even more of the cost is transferred to uplift.

The insufficiency of revenues obtained by offline peaking units would not be a problem if either (1) the cost-efficient provision of first and second contingency reserves were obtained by committing additional units, or (2) the scarcity value of offline resources (especially their value

as substitutes for additional unit commitments) were fully reflected in DA and RT settlements. In case (1) efficiency would not require offline reserves from ICUs, and in case (2) the market prices would attract sufficient ICUs. These conditions do not exist in New England. In particular, the heavy cost of subsidizing additional resources and inadequate compensation for peakers has led to initiatives to procure offline resources in the forward reserve market as well as through "gap RFPs" to solve locational problems.[7]

4.5.2 The FRM solution

In January 2004, the ISO introduced a bid-based Forward Reserve Market. Each unit bids a price ($/MW-month) to provide offline reserves over a six-month period. Each winner is paid the clearing price in the Forward Reserve Market and is then obligated during the ensuing months to bid its reserve quantity into the day-ahead energy market at a price at or above a floor price. The floor price is sufficiently high that there is only a small chance that the unit will be called for energy. There are penalties both for failure to reserve (i.e., failure to bid at or above the floor price) and for failure to start when called. In addition, forward reserve resources that can provide reserve only from an online state are ineligible to receive reimbursement of unrecovered startup and no-load costs. This loss in compensation and any anticipated penalties must be built into the reserve bid.

The initial experience with the FRM is promising. In the first auction for Winter 2004, some 3,474 MW of resources competed to supply the 1,876 MW requirement. Both the ten-minute and thirty-minute products cleared at $4,495/MW-month, roughly 70% of the carrying cost of a new combustion turbine. The two products cleared at the same price because ten-minute was in greater supply, so ten-minute product was used to fill the thirty-minute requirement—an example of price and quantity cascading.[8]

The auction appeared to be successful at selecting units with low opportunity cost for energy (high energy cost) and low performance costs (more reliable units). The winning resources by type are given in Table 3. The lowest average bids came from hydro resources, presumably

[7] See ISO-NE's FERC filing on the gap RFP.
http://www.iso-ne.com/FERC/filings/Other_ISO/Motion_to_Intervene_GAP_01_12_04.pdf
[8] Details of the initial auction are in the 13 January 2004 filing "Compliance Report of ISO New England Inc., FERC Docket No. ER03-1318."

because of their high start reliability and hence low performance penalties. ICUs provided the majority of the reserves.

Table 3. Total megawatt awards by resource type in Winter 2004

Type	Megawatts	Share
ICU	1,113	59%
Hydro with pondage	640	34%
Top end of coal units	63	3%
Duct firing on combined cycles	60	3%
Total	1,876	100%

The forward market was conceived as a means of obtaining offline reserves at prices that were high enough to cover the long-run average costs of fast-response peakers such as ICUs, while low enough to be less costly for the ISO than committing other resources in the absence of ample ICUs. The forward market has apparently been successful in meeting this goal, keeping prices within the window between the costs of ICUs and unit commitment costs. The success of the forward market can be attributed to the fact that intermediate resources did not have incentives to compete materially for these contracts. A likely reason intermediate units did not participate is that their energy opportunity costs of reserves is potentially high. Units frequently scheduled for online duty would be reluctant to reserve blocks of capacity for a six month duration that might otherwise be more profitable when scheduled daily for energy generation.

4.5.3 Need for the FRM remains

The re-design of the reserve markets does not eliminate the motive for a forward reserve market. Under joint optimization of energy and reserves in the DA market, the RT market will still constrain reserve offers to zero. In addition, the energy price cap would continue to limit the revenue that a peaking unit can obtain from being called in periods of shortage.

Another reason that the re-design of the reserve markets does not eliminate the motive for a forward reserve market stems from the binary nature of the unit commitment optimization—units are either committed or not. As a result, the unit commitment optimization must rely on some form of integer programming to meet the constraints of energy and reserves in the DA market. Inevitably, the marginal value of scarce resources—as represented by shadow prices on integer constraints—is not represented, or not accurately represented in this optimization. Shadow prices, measuring the marginal value of a scarce resource, are only well defined when the

decision variables in the constraint are continuous, rather than lumpy. However, unit commitment is necessarily lumpy, since units are either committed or not. Moreover, ISO-NE intends to continue the practice of uplifting unrecovered unit commitment costs.

Thus, the difficulty of capturing the true unit commitment cost of reserves will remain. One can expect, therefore, that the problems will remain, and the forward reserve market may be an integral part of the solution.

4.5.4 Financial vs. physical forward reserves

The prospects for the continued success of the forward market might be bleak if it settles against a DA or RT spot price. As mentioned above, the current forward market does not attract competition from regularly scheduled units because the reserve obligation is physical and cannot be traded. In the re-design, the forward market remains physical, although trading is allowed before the day-ahead market. If instead the market was purely financial, then all suppliers would likely compete for the reserve contracts; moreover, the price in the forward market would likely be the expected price of reserves in the DA and RT markets. Thus, the price in the forward market would reflect the under-pricing of reserves in the DA and RT markets, and the supply of offline reserves from fast-response peakers like ICUs would be deficient.

A forward market for reserves that is basically financial will produce prices that are the expectation of DA prices for reserves, and hence under-value offline resources to the same extent that the DA market does. Therefore, there may again be inadequate installation of ICUs, resulting in too much reliance on the commitment of more expensive thermal units to online duty. Part of the difficulty in resolving this dilemma is that unit commitment entails startup and no-load costs, and minimum generation levels, that make it impossible to accurately estimate the marginal value of reserves. That is, the optimization provides no accurate comparison of the value of an additional 1 MW of offline capacity as compared to the setup costs of bringing online a 100 MW unit with a 20 MW economic minimum (and up to 80 MW available for reserves, including spin but also all categories of offline reserves).

An alternative to avoid this problem is to settle the market before the day ahead at bilateral prices. This is accomplished by requiring that at the time of the day-ahead market a physical resource with a forward reserve obligation must self-schedule to run at least at its economic minimum (EcoMin) throughout the day and must submit an energy bid for its reserve obligation

at the floor price or above. Furthermore, the reserve unit is ineligible to receive any make-whole payment for commitment costs, and is subject to performance penalties for failure to run throughout the day. In this way, the online unit competes with the offline unit on an equal basis. Most importantly, the commitment costs of an online unit appear in its bid to supply forward reserves. Thus, the approach does not presume the desired mix of online and offline resources for offline reserves, but rather enables both types of units to compete on an equal basis to provide these reserves.

This alternative solves the basic problem of how to accurately reward offline resources for the commitment cost savings they provide. The online resource will put its commitment and performance costs in its bid and will only be selected to provide forward reserves if it can do so more economically than an offline resource.[9]

[9] One concern is that spinning resources may receive too little compensation. Reserve compensation for those providing spin (rather than offline reserve) is limited to the energy opportunity cost, which is zero for those not backed down. These units, however, do not incur the costs of those providing offline reserves: (1) there is no energy opportunity cost from bidding at the floor price for the reserve quantity, (2) there is no performance penalty for failure to start and run throughout the day, and (3) there is no loss of the make-whole commitment costs. An alternative would be to expand the forward reserves to include spin; however, the determination of efficient spinning resources varies every day and throughout the day.

5 Discussion of key issues in the proposal

5.1 Absence of a day-ahead reserve market

The proposal includes real-time shortage pricing, but no day-ahead market for reserves. Since marginal costs for reserves are near zero both day ahead and in real time, there is little economic advantage to a day-ahead market. Day-ahead reserve bidding would add complexity for both participants and the ISO.

5.2 Shortage pricing

The proposal includes real-time shortage pricing in the event that there is a shortage of reserves. The prices increase as the shortage extends to higher quality reserves.[10] For example, the reserve prices in a shortage might be set as in Table 4.

Table 4. Hypothetical reserve prices in a shortage

Shortage of	Reserve Price
Thirty-minute operating reserve	$200/MWh
Ten-minute non-spinning reserve	$740/MWh
Ten-minute spinning reserve	$800/MWh

These prices would cascade up to the higher quality reserve products and would be paid to all energy and reserve resources. Thus, if the marginal energy resource was bid at $100/MWh, all energy resources would receive 100 + 800 = $900/MWh during a shortage of spinning reserve.

Shortage pricing is defended on the grounds that it corrects a market failure: the absence of demand response. When supply shortages occur in typical commodity markets, demand sets the price. The price increases until a sufficient number of buyers has reduced demand so that supply and demand are again in balance. This response is largely missing in current electricity markets, including the New England market. Shortage pricing, then, is an attempt to correct the market failure by setting an administrative price in shortages at levels that reflect the marginal value of the short product to demanders.

In theory, shortage prices can be calculated from complex probabilistic models as the marginal value of additional reserves. This approach has proven difficult in practice. A

[10] A theoretical justification for positive reserve payments when reserves are short is developed by Paul Joskow and Jean Tirole, "Reliability and Competitive Electricity Markets," MIT, March, 2004. Besides a scarcity value of reserve capacity, the formula they derive includes an additional term whenever there is risk of grid collapse.

pragmatic alternative, adopted by New York and proposed in New England, is to base the price on decisions operators would have made if additional resources could be acquired through imports. The logic is that if an operator would accept a $1000/MWh import to free-up additional ten-minute spin in a shortage when the energy price is $200/MWh, then spinning reserves implicitly must be worth $800/MWh. Even though the import is not available, the operator's hypothetical choice, based on the $1000/MWh energy price cap, implicitly values reserves.

As of 1 July 2003, New England has implemented a method of shortage pricing for reserves, which is closely related to the New York approach. In particular, when there is a shortage of ten-minute reserves (or a shortage of thirty-minute operating reserves for a period of four hours or more), then the energy price is set at $1000/MWh. Each reserve unit receives its energy opportunity cost, which for example, would be $800/MWh if the unit bid $200/MWh. The impact of the proposal then would be (1) to expand the existing shortage pricing to shortages of thirty-minute operating reserves of short duration and (2) to establish a new methodology for pricing during shortages.

Shortage pricing and capacity

Should we think of the shortage pricing as a reward to capacity or reserves? Since both energy and reserve resources receive the shortage bonus, it is primarily a reward to operable capacity. An inflexible nuclear unit receives the bonus for every MW that it is generating. Likewise, an offline peaker receives the bonus for its fast-response electricity. The only resources that do not receive the bonus are resources that are declared inoperable or are unable to come online during the period of the shortage. For shortages that are caused by unanticipated real-time contingencies, there may be a few operable resources that are not entitled to the bonus because of inflexibility (start times, minimum down times, or ramping limits). For the most part, all operable capacity receives the shortage bonus. This is economically efficient, since all resources providing energy or reserves during a shortage are contributing to reliability.

Shortage pricing does have several advantages over a traditional ICAP market as a mechanism for paying for capacity. First and most important, the high spot prices motivate both supply and demand response to address the shortage. Without high spot prices at times of shortage, supply and demand response will remain undeveloped. Second, only capacity that is available in times of shortage is paid. Inoperable capacity goes unrewarded. In contrast,

traditional ICAP rewards capacity based on average availability (in contrast, New England's proposed LICAP market avoids this serious flaw). Thus, a gas-fired unit could buy non-firm gas and receive nearly full ICAP payments, since the unit is nearly always available. However, the unit is likely to be unavailable when it is most needed—during a gas shortage. An extreme example of this was the "cold snap" of January 2004.[11] Third, to the extent that shortages are unanticipated day-ahead, resources offering greater flexibility are better able to benefit from the shortage pricing. Faster starting units that can ramp quickly are better able to capture high spot prices during unanticipated shortages.

A potential disadvantage of shortage pricing is that it may create incentives for suppliers to create a shortage in order to benefit from the high spot prices on their remaining operable capacity. This problem is addressed in the definition of the proposed LICAP product. Since the shortage price is deducted for all LICAP resources, regardless of whether the resource is providing energy or reserves, there is no gain from creating a shortage. The LICAP resource can avoid a loss by providing energy or reserves during shortages, but cannot gain by creating shortages. Since LICAP is purchased based on peak energy plus reserve needs, load is fully hedged from the shortage prices and real-time supplier market power is mitigated effectively.

Once the LICAP market is introduced, the penalty factors will perform an important role: to send a high real-time price signal to motivate suppliers to increase supplies and demanders to decrease demands. Thus, the shortage prices, together with the definition of the LICAP product, provide the essential real-time incentives for suppliers to increase supply and demanders to reduce demand in times of shortage.

The shortage prices do not play a direct role in determining the level of new investment. Rather, the LICAP market will fill this important role. The reason is that, since the shortage prices are deducted for all LICAP resources, the prices cannot motivate new investment. Instead, the shortage prices, together with lost LICAP payments, serve as a penalty for resources that are unavailable during shortages.

The proposed penalty factors are set to be consistent with actions that system operators would take to restore reserves if such offers were available. This pragmatic approach does

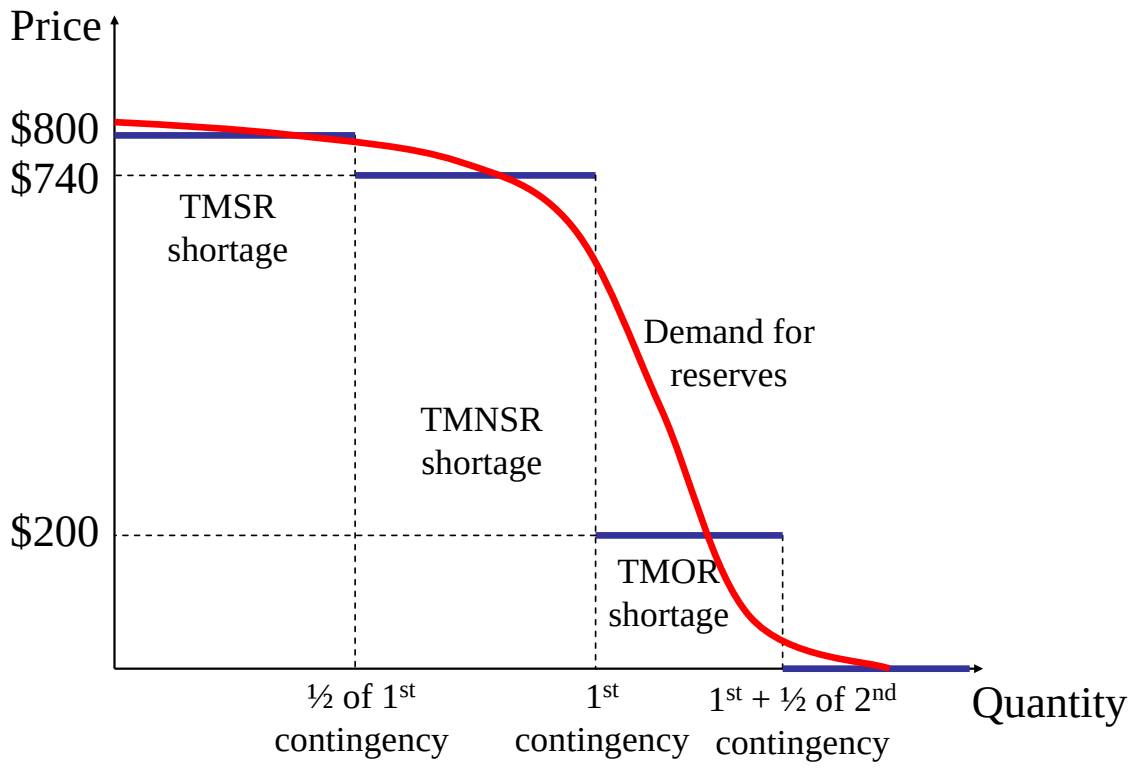
[11] "Interim Report on January 14-16, 2004 Cold Snap" at http://www.iso-ne.com/special_studies/.

rationalize the choices system operators make given the energy price cap and NERC and NPCC reliability rules.

5.2.1 Shortage pricing in the real-time dispatch objective function

One question that must be resolved is whether shortage pricing is included in the objective of the real-time dispatch optimization. This is the natural approach if we treat shortage pricing in the context of a reserve demand curve, and it is the approach taken in our mathematical formulation of the reserve proposal, contained in Appendix 1. With this interpretation, the shortage price represents load's maximum willingness to pay for reserves. Hence, if the TMOR shortage price is $200/MWh, then the optimization would fail to maintain TMOR reserves when the cost exceeded $200/MWh. Thus, by including shortage pricing in the objective, the optimization will satisfy reserves only when it is economic do to so.

Figure 1. Reserve demand curve under shortage pricing



Thus far, NERC and system operators have been unwilling to adopt this economic approach. Operators do not base the level of reliability on costs and benefits, but rather on reliability constraints that must be satisfied, *regardless* of the cost. Under this traditional view, shortages

only arise when it is physically impossible to satisfy the reserve constraints. Then the shortage prices do not represent a demand curve, but rather bonuses that are paid to those providing energy or reserves during shortages. This is perhaps the better interpretation, since shortage pricing is an implausible representation of the true demand for reserves. In particular, it yields a staircase demand curve with dramatic discontinuities at the reserve requirements, as given by the blue step-function of Figure 1. Any model of the value of reliability would value the first MW of TMOR more than the last (at the requirement), and the first MW of TMOR beyond the requirement would not have a value of 0. Hence, the red curve is likely a better representation of the demand for reserves.

Regardless of the approach taken, it is important that the real-time market operations be consistent with actual system operations. Inconsistencies between market and system operations create incentives for gaming by participants.

Both the step function and the curve in Figure 1 have two important features. First, the price for a shortage of ten-minute reserves is much higher than for a shortage of thirty-minute operating reserves. The reason for this is that a ten-minute reserve shortage implies that load would be shed in response to the first contingency; whereas the thirty-minute operating reserve shortage implies that load would be shed if both the largest and second-largest resources failed, which has much lower likelihood. Second, the shortage price for both ten-minute spin and non-spin are roughly equal, since ten-minute non-spinning reserve can substitute for spinning reserve simply by turning on. If the shortage prices for ten-minute spin and non-spin differed by more than the start-up costs of the non-spin resource, then the resource would have an incentive to start in order to receive the higher spinning reserve price, regardless of the resource's dispatch instructions.

5.2.2 Shortage pricing should be included in the real-time energy price

There are two seemingly equivalent ways to implement shortage pricing: (1) include the shortage price in the real-time energy price, or (2) add the shortage price as a bonus in the real-time settlement. As an example, consider a unit providing 1 MW of energy in real time, when the energy price is $60, absent shortage pricing, and the expected shortage price is $20. Then in (1) the unit receives the revised real-time energy price of 60 + 20 = $80; whereas in (2) the unit receives the real-time energy price of $60 plus a bonus of $20 for a total of $80. Although the

real-time compensation is the same with either approach, the two approaches imply different bidding in the day-ahead energy market.

If the shortage price is included in the real-time energy price, the day-ahead price will reflect the expected shortage price. This is desirable, since then the day-ahead commitment will reflect the greater scarcity of resources, as implied by the expected shortage price. Notice that a supplier with a marginal cost of $30, anticipating a $20 shortage price, would offer (absent market power) $30 in the day-ahead market. Most importantly, the day-ahead offer would not include the $20 opportunity cost of providing energy. The reason is that the supplier has a better way of expressing the $20 opportunity cost: the supplier can submit a DEC bid at $50. This virtual bid has the supply "buy back" its energy whenever the day-ahead clearing price is below $50, which effectively shifts the sale to the real-time market. Notice that the supplier is committed based on the physical day-ahead offer of $30 (the resources marginal cost); hence, the day-ahead schedule is efficient. However, the virtual bid appropriately represents the resources expected opportunity cost of selling energy day-ahead, rather than selling in real-time. A further advantage of using the virtual bids to express the opportunity cost of providing energy is that a different DEC bid can be given in each hour. In contrast, the physical offer applies throughout the day. The added flexibility of the virtual bids is important, since the expected shortage price (and hence the opportunity cost of providing energy) will vary by hour—peak hours have higher expected shortage prices.

If instead the shortage price is treated as a separate capacity adder that is paid in settlement to all those providing energy or reserves, rather than adding the shortage component to the real-time energy price, then the day-ahead energy price will not reflect the scarcity implied by the expected shortage price. The problem with this is that the ISO will schedule too few resources, based on the lower day-ahead price. This is inefficient, since the day-ahead commitment should reflect the possibility of shortage in real time. For this reason, it is best to include the shortage price in the real-time energy price.

5.3 Forward reserve market

The proposal includes a forward market for offline reserves. This improves upon the existing forward reserve market by (1) allowing portfolio offers, (2) allowing bilateral trade, and (3) adding locational requirements.

We support retaining and improving the forward reserve market. Our reasons are several.

- A long-term product is a better instrument for valuing emergency energy, and provides a stronger basis for efficient investment and performance. With the purchase sufficiently forward, the fixed costs of offline units become variable. Resources, especially those on the margin, can enter or exit, so the clearing price is sufficient to cover the total cost of the marginal unit.
- The revenues are directly targeted to physical resources providing fast-response reserves.
- Participants are able to plan ahead knowing their obligations well in advance.

Portfolio offers and bilateral trade in the forward reserve market

By allowing portfolio offers and bilateral trade up until the day-ahead market, the forward reserve market gives participants the flexibility to make adjustments to improve efficiency. For example, a hydro unit's cost of providing reserves depends on the opportunity cost of using the water for energy. This opportunity cost is highly variable. When storage is limited and the energy price is high, the owner of a hydro unit may prefer to supply energy, and provide reserves with an alternative unit. Likewise, the commitment costs of an online unit are dependent on energy prices throughout the day. Portfolio offers allow a supplier to satisfy its obligation at least cost.

Bilateral trade of the forward reserve product enables suppliers to make adjustments in light of changed circumstances. Suppliers who are short of their obligation on a particular day are able to buy supply before the day-ahead market. However, the forward market does not settle against the DA or RT reserve market. This prevents the forward reserve price from being undermined by low spot prices that omit costs already sunk.

Incentives for self-selection in the forward reserve market

The forward reserve market creates incentives for self-selection in three dimensions: (1) low energy opportunity cost (that is, a high energy cost), (2) high reliability in providing emergency energy, and (3) for online units, low commitment costs. The first is accomplished by the requirement to bid forward reserve capacity into the energy market at a price at or above a floor price. The second comes from penalties for nonperformance (either a failure to reserve capacity or a failure to supply energy when called). The third comes from eliminating the make-whole

commitment costs to online generators providing forward reserves. An online resource reserves its capacity by self-scheduling throughout the day and by bidding its reserved capacity at the floor price or above. In this way, online resources can compete on an equal basis with offline resources to provide offline reserves.

Compensation for resources held in reserve

Another issue is how to compensate resources with a forward reserve obligation that are held back rather than dispatched for energy. This practice of holding units back out of merit is rare in New England. The issue is whether energy opportunity costs should be paid on an individual basis (pay-as-bid) or whether a market-clearing energy opportunity cost should be paid (uniform price). For two reasons, we favor the pay-as-bid approach for those with an obligation in the forward reserve market. First, to the extent that uniform pricing results in extra spot compensation, the equilibrium forward reserve price will fall. Hence, pay-as-bid pricing reduces generator risk by putting more of the compensation in the forward reserve price (a constant payment) rather than the volatile market-clearing energy opportunity cost. Second, pay-as-bid pricing reduces incentives to distort the energy bid to increase the energy opportunity cost payment. With uniform pricing, a large supplier of reserves may have a strong incentive to reduce its energy bid below marginal cost if the supplier believes it will be held back for reserves.

Another reason for paying as-bid is that it simplifies the optimization. Under simultaneous optimization of reserve and energy, the commitment of an additional unit must weigh the operating costs against startup and no-load costs.

A primary purpose of the simultaneous optimization of energy and reserves is to efficiently commit additional units to make sufficient reserves available. Additional generation is added only if its commitment cost is less than the cost of backing units down. For units providing forward reserves, this tradeoff is internalized by suppliers through their offers to supply forward reserves. For other units, the tradeoff is made in the joint optimization of energy and reserves. Commitment costs and energy opportunity costs are treated in a symmetric way under the pay-as-bid rule.

5.3.1 Locational reserve requirements

The vast majority of operating reserve credits (uplift) are the result of commitments made after the day-ahead market to satisfy reliability constraints in either Boston or Connecticut.[12] Supplemental commitments of non-economic resources in load pockets have the impact of reducing the energy prices in the constrained areas. This discourages investment in areas where it is most needed. The proposal addresses this problem by including locational reserve constraints in both the forward reserve market and in shortage pricing.

Since the locational constraints are all thirty-minute constraints associated with the second contingency, they are addressed by an additional TMOR product in each import constrained area. This product is procured in the forward reserve market based on peak requirements. The price for this product would be at least as high as the system-wide product, since the local product can substitute for the system-wide product.

A potential problem with this approach is that it over-supplies reserves in load pockets in off-peak periods. This happens because the requirement for local reserves depends on whether the local market is import constrained. In contrast, system-wide reserves, which do not vary with system conditions, are not over-supplied. In theory, one could avoid this problem by defining the local reserve product as an option. The ISO then calls on the required quantity to provide local reserves according to a day-ahead schedule. However, we doubt that the added complexity would yield much in the way of added benefits. The simple approach is efficient if local operating reserves are most efficiently provided by offline units that would not change their behavior in the day-ahead and real-time markets based on knowing the hours they are providing reserves. This seems plausible. Appropriate adjustment of availability penalties can further address this issue by, for example, lowering penalties during the off-peak period.

Reliance on the forward market to provide local reserves is sensible. Market power is much more of an issue in constrained local markets. And forward purchase better mitigates these market power problems. Attempts to include out-of-merit commitment costs in a local reserve spot price are problematic at best. Mathematically, it is not possible to determine the “marginal” commitment cost, since unit commitment is a lumpy decision.

[12] See Patton, David B. et al., “Six-Month Review of SMD Electricity Markets in New England,” February 2004.

5.3.2 Performance incentives

The proposal recognizes the important role of performance penalties in the forward reserve market. These penalties motivate efficient generator decisions. Although the penalties have yet to be finalized, basic principles for setting the penalties have been established. We review these here.

Basic Principles

The penalties are of two types: availability and performance. Availability measures the megawatts of reserve available either day-ahead or real-time. Performance measures the success of the resource to respond as promised in ten or thirty minutes. The penalties should satisfy the following principles.

Allow efficient breach

The size of the penalty should depend on the harm caused, which varies with system conditions and the type of violation. Penalties should be larger when energy prices are higher, since this indicates tighter supply conditions and a greater value of reserves. Penalties approximating the harm from breach encourage efficient breach by suppliers. The penalty makes the buyer (the ISO on behalf of load) whole, and induces the supplier to breach and pay the penalty, whenever the cost of supplying exceeds the harm caused to the buyer.

Encourage early reporting of unavailable resources

It becomes more difficult for the ISO to make adjustments as we get closer to real-time. Hence, an unavailable resource should be encouraged to report its condition in the day-ahead market, rather than the real-time market. Since reserve units are rarely called, the performance penalty needs to be substantially above the availability penalty to discourage an unavailable resource from masquerading as available. Thus, other things being equal, the penalties are ordered as follows:

DA Availability Penalty < RT Availability Penalty << Performance Penalty

Induce energy offers above the floor price

A unit with energy costs below the energy price may find it profitable to bid its energy costs, rather than the energy floor price. A penalty that took away its energy profits would prevent this.

Prevent resources from leaning on pool commitment

Unavailable resources should be encouraged to use the bilateral market to satisfy their obligation, rather than rely on the pool commitment, in which commitment costs are uplifted to all load. The availability penalty needs to encourage the use of the bilateral market.

Discourage market power in the bilateral market

Large penalties enhance the ability of sellers in the bilateral market to exercise market power, since the penalty is a ceiling on the bilateral price. Hence, the penalties should be set as low as possible and still maintain proper incentives. Lower penalties also reduce the risk of providing reserves, and thus reduce the cost of reserves.

Encourage an efficient mix of resources

The penalties should encourage an efficient mix of the two technologies for reserves: fast-start units and the commitment of online resources at their economic minimum. The two compete on an equal basis before the day-ahead market, but once committed, the marginal cost of providing reserves is zero. Hence, the penalties should induce suppliers to cover their obligations before the day-ahead market. Then the online resources will include their foregone commitment costs in their offers to supply reserves.

Discussion

The basic principles outlined above appear both sound and comprehensive. Nonetheless, setting efficient penalties is complex. Often there are countervailing forces, which necessitate a balance. For example, since reserve units are rarely called, the performance penalty needs to be much higher than the availability penalty, so that unavailable units do not pretend to be available. But at the same time the spread in penalties cannot be so large so as to induce an available, but somewhat unreliable, offline resource from declaring unavailability as soon as it looks likely that it will be called to provide energy, and hence be at risk of the high performance penalty. As the likelihood of being called increases, the availability penalty must increase as well.

Given the importance of the penalties in motivating efficient supplier decisions, we recommend that proposed penalties be thoroughly studied and tested.

6 Recommendations

The proposed market design has changed substantially during the period of our review. Many of our recommendations are addressed in the current proposal. Some of our conclusions deserve further analysis, testing, and refinement. The appendix outlines an approach to testing using agent-based simulation.

The core elements of the proposal are: (1) an annual (or semiannual) forward reserve market for offline reserves, (2) joint optimization of energy and reserve constraints both day ahead and in real time, and (3) a real-time operable capacity bonus based on penalty factors during reserve shortages.

6.1 The forward reserve market for offline reserves plays an important role

The forward reserve market is an effective way to motivate efficient investment in reserve resources. Pricing offline reserves is done much better on a forward basis, before the costs of providing the service have been sunk. Also long-term prices provide a better basis for investment than volatile spot prices. Bilateral trading and portfolio offers enable offline and online resources to better compete for reserve supply, improving the efficiency of the mix of reserve resources. Locational reserve constraints and associated local reserve products assure that reserve resources in import constrained zones receive sufficient compensation to motivate the reserve investment where it is needed most. Finally, performance incentives are addressed through a system of performance penalties. Under the proposed structure, the forward reserve market identifies those that can best provide reserves. Self-selection occurs on three dimensions: (1) resources with low energy opportunity costs, (2) more reliable resources, and (3) online resources with low commitment costs.

6.2 No day-ahead reserve availability bids

There is little justification for day-ahead availability bids when reserves are defined as idle capacity, capable of producing energy in ten or thirty minutes. Since commitment costs are paid, marginal costs of reserve, other than energy opportunity costs, are zero for all reserve resources both day-ahead and in real time. In this case positive availability bids only can serve to distort the schedule and dispatch away from an efficient outcome. Moreover, absent market power, day-ahead availability bids would not provide much extra revenue for flexible resources, since we can expect the day-ahead price to equal the expected real-time price, which is based on bids set

to zero. The elimination of day-ahead availability bids greatly simplifies the market and eliminates any exercise of market power day-ahead.

6.3 Adopt the performance-based LICAP product to hedge the shortage price

Shortage pricing is a reward for operable capacity in times of shortage. Without the proposed LICAP market, shortage pricing would introduce incentives to create real-time shortages, and thus would undermine reliability. This problem is fully addressed by the proposed LICAP market. Shortage prices are deducted from the payments to LICAP resources. Thus, LICAP resources have no incentive to create shortages, and load is fully hedged from shortage prices.

6.4 Load should not participate directly in the procurement of reserves

Reserves are a reliability product and a public good in current electricity markets—if given a choice an individual load would free-ride on the reserves (and reliability) provided by others. As such, reserves are best purchased by the ISO. Load's participation should be in supplying reserves through dispatchable load, not in deciding when and how much reserves to purchase.

6.5 LICAP and forward reserve markets are complementary

Both the LFRM and LICAP markets are important in motivating efficient investment in capacity of the right type and in the right location. The LICAP market provides the foundation for capacity investment in interaction with the spot energy and reserve markets. LFRM provides supplemental compensation where needed to flexible resources. Over time, LICAP should provide the vast majority of capacity compensation to resources, regardless of whether the capacity is used to provide energy or reserves. However, LFRM will remain an important market to price flexible capacity at times and in locations where such capacity is scarce.

Appendices

The following appendices address several issues in greater detail. Appendix 1 provides a mathematical formulation of the proposed markets. Appendix 2 reviews the results of testing the design with agent-based simulation. Appendix 3 outlines how settlements work in a co-optimized market for energy and reserves. Finally, appendix 4 considers the implications of including out-of-merit commitment costs in the reserve price.

Appendix 1: A mathematical formulation of the market design

This appendix describes a mathematical formulation for the market clearing and settlement rules in energy and reserve markets that represent ISO-NE's proposed reserve market design. Our objective is to present the proposed market rules with such a clarity that facilitates the complicated tasks of design, analysis, simulation and communication within a sophisticated multidisciplinary group.

First, in section 1, we summarize the notation. Then, in sections 2-4, we describe the market clearing models for forward reserve, day-ahead and real-time markets. Lastly, in section 5, we describe the settlement rules for the different products: forward reserve, financial transmission rights, energy and operating reserve.

1. Notation:

Tables A1-A3 present the notation and definitions.

Table A1. Indices and sets

Symbol	Definition
j:	Generator $(j = 1, 2, ..., J)$
n:	Bus node or zone $(n = 1, ..., N)$.
G_n	Generator set at node n; $G_0 \equiv \bigcup_{n=1}^{N} G_n = \{1, ..., J\}$.
k:	Contingency state $(k = 0, 1, ..., K)$; 0 indicates the normal state.
l:	Transmission line $(l = 1, 2, ..., L)$
i, m:	Reserve type (1=TMSR; 2=TMNSR; 3=TMOR + Replacement Reserve)
t:	Time $(t = 1, 2, ..., T)$
s:	Segment on a bid curve $(s = 1, ..., S)$

Table A2. Variables

Symbol	Definition
P:	Generation output
Q:	Forward reserve (to meet 100% of the expected peak requirements for TMNSR and TMOR requirements plus extra Replacement Reserves)
R:	Operating Reserve
S_m:	Reserve shortage (in violation of the reserve requirement; $S_1 \equiv 0$)
D	Demand bid (dispatchable)
u:	Unit commitment decision (=0 or 1)
X_l:	Power flow on line l
p_t^{DA}:	Day-ahead energy price at time t
p_t^{RT}:	Real-time energy price at time t
p_m^{FR}:	Forward price of type m reserve

Table A3. Parameters and functions

Symbol	Definition
D^k:	Energy demand in state k
E	Energy imports
R_{mn}^{req}:	Reserve requirement for type *m* at node *n* (system requirement, when *n*=0)
π_k:	Probability of state *k*
β_{nl}^k:	Power transfer distribution factor on line l for injection at node *n* in state *k*
$EcoMax$:	Economic maximum generation capacity
$EcoMin$:	Economic minimum generation capacity
$R^{\max}$	Maximum ramping capability available for operating reserve (day-ahead)
$X^{\max}$	Transmission line capacity
$\Delta^{up}, \Delta^{down}$	Ramping rate limits (for upward and downward ramping)
c_m^{FR}	Minimum energy bid for forward reserve resource of type *m*
(r_{jm}, Q_{jm})	Price and quantity of forward reserve bid for generator *j* and reserve type *m*
(c_{js}, q_{js})	Price and quantity of energy bid for generator *j* and segment s
$C_j(P_j)$:	Energy cost function (based on bids) of generator *j* at output level P_j, which includes supply bids from the forward reserve resources $= \underset{x}{Min}\left\{\sum_{s=1}^{S} c_{js}x_{js} \middle\| \sum_{s=1}^{S} x_{js} = P_j; 0 \le x_{js} \le q_{js}\right\}$
(b_{ns}, q_{ns})	Price and quantity of energy demand bid for segment s at node *n*
$B_n(D_{nt})$	Benefit function based on energy demand bids $= \underset{x}{Max}\left\{\sum_{s=1}^{S} b_{ns}x_{ns} \middle\| \sum_{s=1}^{S} x_{ns} = D_{nt}; 0 \le x_{ns} \le q_{ns}\right\}$
$H_m(S_{mt}^k)$	Shortage cost of operating reserve based on administratively set penalty factors
f_{0j}:	Start up cost
f_{1j}:	No load cost

Current reserve requirements are as follows:

Ten-minute spinning reserve (TMSR):	½ of first contingency.
Ten-minute non-spinning reserve (TMNSR):	½ of first contingency.
Thirty-minute operating reserve (TMOR):	½ of second contingency.
Thirty-minute replacement reserve:	¼ of second contingency.[13]

2. Forward reserve market

The forward reserve market allocates by auction the capacity credits for operating reserves, including extra 30-minute replacement reserves, but spinning reserves are acquired in the real-time market. The auction is cleared under the condition of minimizing the total bid cost with a simple linear program model:

$$\underset{x_{jm}}{Min} \sum_{j\in 1}^{J}\sum_{m=2}^{M} x_{jm} r_{jm} \tag{2.1}$$

where x_{jm} is the quantity of reserve type m provided by resource j. The system and local reserve requirements are described in (2.2), which incorporates forward cascade by allowing any excess of a superior reserve (e.g. TMNSR) to substitute for an inferior reserve (e.g. TMOR).

$$\sum_{j\in G_n}\sum_{i=2}^{m} x_{ji} \geq \sum_{i=2}^{m} R_{in}^{req},\ n=0,...,N;\ m=2,...,M \tag{2.2}$$

$$0 \leq x_{jm} \leq Q_{jm},\ j=1,...,J;\ m=1,...,M \tag{2.3}$$

3. Day-ahead market

The day-ahead market clearing is based on a standard formulation of a unit commitment model that captures the essential features of energy and reserve markets. The model includes both

[13] There is no separate market for replacement reserves, which will be an extension of TMOR.

demand and supply bids. The objective is to maximize the expected total social surplus, which equals the total consumer benefit (based on the information in demand bids) minus operating energy cost, $C_j(P_{jt}^k)$, no-load cost, $f_{0j}u_{jt}^k(1-u_{j,t-1}^k)$, and start-up cost, $f_{1j}u_{jt}^k$.

$$\underset{u,D^0,P,R,q,X}{\text{Maximize}} \sum_{\text{k}=0}^{\text{K}}\sum_{t=1}^{\text{T}}\sum_{j=1}^{J}\sum_{n=1}^{N}\pi_k\left[B_{nt}(D_{nt})-C_j(P_{jt}^k)-f_{0j}u_{jt}^k(1-u_{j,t-1}^k)-f_{1j}u_{jt}^k\right] \qquad (3.1)$$

In (3.2)-(3.3), we specify the net injection at each node and the power flow on each line. The injection at each node equals the total generation output and imports net of the load, including demand bids, at the node.

$$q_{nt}^k=\sum_{j\in G_n}P_{jt}^k-D_{nt}-D_{nt}^k+E_{nt}^k,\ \text{for n}=1,...,N;\ t=1,...,T;\ k=0,...,K \qquad (3.2)$$

The power flow on each line is derived from the injection at each node multiplied by the power transfer distribution factor over the line.

$$X_{lt}^k=\sum_{n=1}^{N}\beta_{nl}^k q_{nt},\ \text{for } l=1,...,L;\ t=1,...,T;\ k=0,...,K \qquad (3.3)$$

In (3.4) and (3.5), we introduce the energy balance requirement and the reserve capacity requirement.

The energy balance requirement states that the net energy injection for the system as a whole must be zero.

$$\sum_{j=1}^{J}q_{jt}^k=0. \qquad (3.4)$$

The system and local reserve capacity requirements can be stated in the following form, which allows for cascading of higher to lower quality reserve types,

$$\sum_{i=1}^{m}\sum_{j\in G_n} R_{ijt}^{k} \geq \sum_{i=1}^{m} R_{int}^{req},\ \text{for } m=1,...,M;\ k=0,...,K;\ t=1,...T;\ n=0,...,N. \qquad (3.5)$$

Constraints (3.6) - (3.9) represent bounds on generation output, ramping capability, transmission line capacity and total capacity. The generation output is bounded by the upper and lower economic operating limits,

$$u_{jt}^{k} EcoMin_j \leq P_{jt}^{k} \leq u_{jt}^{k} EcoMax_j \qquad (3.6)$$

The reserve capacity is limited by the ramping rate of the generator,

$$0 \leq R_{mjt}^{k} \leq R_{mjt}^{\max} \qquad (3.7)$$

The power flow on each line must satisfy the transmission line capacity constraints

$$-X_l^{\max} \leq X_{lt}^{k} \leq X_l^{\max} \qquad (3.8)$$

The total unit capacity committed to energy and reserve must stay within the generation capacity limit,

$$P_{jt}^{k} + \sum_{m=1}^{M} R_{mjt}^{k} \leq EcoMax_j \qquad (3.9)$$

The security constraint ties the above formulation together, for the model can otherwise be decoupled into K+1 separate sub-models. The system security constraint (3.10) requires that the total unit capacity committed for each generator stays the same for all contingencies.

$$P_{jt}^{k} + \sum_{m=1}^{M} R_{mjt}^{k} = P_{jt}^{0} + \sum_{m=1}^{M} R_{mjt}^{0}, \text{ for } k = 0,...,K;\ \mathrm{j} = 1,...,J;\ t = 1,...,T \qquad (3.10)$$

4. Real-time market

The real-time market clearing depends on a dispatch model which differs from the unit commitment model in three key aspects: 1) the commitment of individual units has been decided; 2) the allocated operating reserves are priced according to a shortage cost function; and 3) the time scale in real-time is finer. The specific differences between the real-time dispatch model and the unit commitment model are summarized below. In the dispatch model,

- the unit commitment decision (*u*) is given.
- the reserve requirement in (4.5) includes the slack variable S_{mn}^{k}
- the demand for reserve is considered in the shortage cost function $H_{mn}(S)$
- the ramping constraint is explicitly represented in (4.10)
- the prevailing state (*k*) is emphasized

The objective of the dispatch model is to maximize the total social surplus, including the energy benefit (for the dispatchable load) minus the energy cost and the shortage cost of reserve,

$$\underset{D,P,R,S,q,X}{\text{Maximize}} \sum_{t=1}^{\mathrm{T}} \sum_{j=1}^{J} \sum_{n=0}^{N} \left[B_{nt}(D_{nt}) - C_{j}(P_{jt}^{k}) - \sum_{m=2}^{M} H_{mn}(S_{mnt}^{k}) \right], \qquad (4.1)$$

subject to the constraints in (4.2) - (4.10). In (4.1), the energy benefit function, $B_{nt}(D_{nt})$, represents the energy demand; the energy cost function, $C_{j}(P_{jt}^{k})$, represents the energy supply; and the shortage cost function, $H_{mn}(S_{mnt}^{k})$, represents a price-sensitive demand function for non-spin or operating reserves. The demand for spinning reserve is based on pure engineering considerations and is not price-elastic.

Equation (4.2) depicts that the net injection at each node equals the total generation output and imports net of the load, including demand bids, at the node.

$$q_{nt}^{k} = \sum_{j \in G_n} P_{jt}^{k} - D_{nt} - D_{nt}^{k} + E_{nt}^{k}, \text{ for n} = 1,...,N;\ t = 1,...,T \tag{4.2}$$

In (4.3), the power flow on each line equals the sum of the product of the injection at each node and the power transfer distribution factor.

$$X_{lt}^{k} = \sum_{j=1}^{J} \beta_{jl}^{k} q_{jt}, \text{ for } l = 0,...,L;\ t = 1,...,T \tag{4.3}$$

In (4.4), the energy conservation principle requires that the net energy injection for the system is zero.

$$\sum_{j=1}^{J} q_{jt}^{k} = 0. \tag{4.4}$$

In (4.5), the system and local reserve capacity requirements are met, allowing for cascading of higher to lower quality reserve types.

$$\sum_{j \in G_n} R_{mjt}^{k} + S_{mn}^{k} \geq R_{mnt}^{req}, \text{ for } m = 1,...M;\ t = 1,...T;\ n = 0,...,N \tag{4.5}$$

The generation output is bounded by the upper and lower economic operating limits,

$$u_{jt}^{k} EcoMin_{j} \leq P_{jt}^{k} \leq u_{jt}^{k} EcoMax_{j} \tag{4.6}$$

The reserve capacity is limited by the ramping rate of the generator,

$$0 \leq R_{mjt}^{k} \leq R_{mjt}^{\max} \tag{4.7}$$

The power flow on each line must satisfy the transmission line capacity constraints

$$-X_{l}^{\max} \leq X_{lt}^{k} \leq X_{l}^{\max} \tag{4.8}$$

The total unit capacity committed to energy and reserve must stay within the generation capacity limit,

$$P_{jt}^{k} + \sum_{m=1}^{M} R_{mjt}^{k} \leq EcoMax_{j} \tag{4.9}$$

The changes in power generation over time are constrained by up- and down-ramping limits,

$$\Delta^{down} \leq P_{jt}^{k} - P_{j,t-1}^{k} \leq \Delta^{up} \tag{4.10}$$

In (4.1), a set of shortage cost functions replace the reserve requirements in (3.5), assuring cascade substitution so that the reserve price increases with quality. It is important that the market implications of the reserve demand function match the results of actual system operation, because any predictable inconsistency between market and system operations creates gaming opportunities. The Dec game in California electricity market offers an extreme example. Constraint (4.10) is relevant in real-time scheduling when load variation may be erratic in a short time interval.

The effect of shortage pricing is to create a positive reserve price for operating reserves when the reserve requirement (e.g. TMOR for which the requirement is relaxed) is violated. This effect will propagate through the forward cascade condition in (4.5) and the coupled energy-reserve constraint in (4.9) resulting in higher energy prices.

5. Settlement

This section briefly summarizes the settlement rules for the following products: forward reserve, energy, FTR and operating reserve.

Forward Reserve (FR)

Each available forward reserve resource receives payments based on a uniform market price p^{FR}. The FR payments are similar to capacity payments in that they are made regardless of whether the system operator decides to use the resource for energy or reserve. A forward reserve resource must bid in the DA and RT markets with a bid price that must be higher than a threshold $c_j \geq c^{FR}$, which is obtained from a formula based on a fixed heat rate and the current fuel price. Forward reserve resources are allowed to receive payments from the energy market, but must forego payment from providing reserve when called upon to provide energy.

Energy

The Energy product is traded both DA and in Real Time. Energy bids and offers are taken in the DA market, and determine the DA settlement. Then the RT market is used to settle deviations from the DA settlement. Adjustments in the quantity of exchanged energy can be provided from spinning and non-spinning resources and changes in adjustable loads. Energy is paid LMPs. The DA energy price is p^{DA}, the shadow price of (3.2), and the RT energy price is p^{RT}, the shadow price of (4.2).

Financial transmission rights (FTR)

The FTRs are paid the nodal LMP differences.

Operating reserve

The spinning reserve is credited in an amount equal to the lost opportunity cost, $p^{RT} - c_j$. The shadow price of spinning reserve in (4.9) equals the marginal lost-opportunity cost but is not used in settlement.

The proposed reserve payment is set administratively at a price equal to the nodal marginal shortage cost, $\sum_{i=m}^{M} H'_{in}(S^k_{int})$, for $m \geq 2$, including TMNSR, TMOR plus Replacement Reserve. The forward reserve resources are excluded from receiving the reserve payment.

Appendix 2: Simulation and testing of the proposed design

We outline a plan for simulation and testing of the proposed design of the new reserve markets in New England. In collaboration with EPRI, ISO-NE is currently implementing this plan. Simulation tests offer two important advantages. First, simulation necessitates specification of the proposed market rules that are sufficiently explicit and precise that they can be programmed—including bidding by participants, market clearing, and settlements. Second, simulation allows the proposed design to be tested and evaluated in a "wind tunnel" without risk of actual market failure. Therefore, the objectives of the simulation study are to verify the completeness and accuracy of the proposed design, then to test it under a variety of "what-if" scenarios (motivated by the issues discussed in our review), and finally to describe and evaluate the predicted performance of the design.

To ensure that the simulations enable realistic evaluation of efficiency and the role of strategic behavior, we recommend agent-based simulation as an important component. Programmed "agents" can mimic participants' attempts to maximize their profits individually from participating in the new markets within the rules and operating procedures of the proposed design.

A primary purpose of the proposed design is to optimize schedules and real-time dispatch to satisfy energy and reserve constraints simultaneously. It does not include bid-based reserve markets, either day ahead or real time, but it does include shortage pricing as a capacity adder in real time. One important question is the impact of alternative penalty factors. Most directly, the penalty factors impact revenues of all operable capacity. The shortage pricing also rewards reliability and flexibility.

One concern is whether shortage pricing introduces an incentive for large suppliers in import-constrained zones to create shortages. The assumption is that market power mitigation prevents the suppliers from economically withholding capacity through high energy offers (at or near the price cap of $1000/MWh). Rather, the large supplier physically withholds sufficient capacity to create a local shortage of TMOR, triggering the shortage price. The incentive to do so depends on the size of the penalty factor, the quantity of unhedged capacity of the large supplier, and the tightness of supply.

Another important analysis focuses on the forward reserve market. An analysis of the impact of alternative penalties is essential. Penalties establish the performance incentives. The goal of the penalties is to induce efficient decisions on the supplier in the operation of the supplier's capacity. The penalties also impact the substitution of online and offline resources in providing offline reserves.

A third issue is the handling of local reserve constraints and the possibility of inefficiently acquiring too many local reserves in off-peak periods, since the quantity of local reserves is based on peak conditions.

The proposal allows portfolio offers and bilateral trade of the forward reserve obligation. However, there is no spot market where a short supplier can purchase reserve capacity to fulfill its forward obligation. Such a spot market could improve short-run efficiency, but it might undermine an important feature of the forward pricing, the inclusion of costs that become sunk on a spot basis. One question is whether such a spot market for reserves would undermine the forward reserve price. This question can be addressed in a number of steps.

- The first step is to simulate the real-time markets for energy and reserves when they are accompanied by a financial forward reserve market. This step assumes a single settlement of all transactions other than the financial forward reserve.
- The second step is to compare the results from the first step with those when a physical forward reserve market precedes the real-time markets. This step assumes separate settlement of the forward reserve and real-time markets.

In these first two steps the markets are simulated without taking account of the day-ahead unit commitment process—for practical reasons, it is useful to simplify assumptions on unit commitment at the beginning of the study and relax them later.

- The third step introduces the day-ahead unit commitment process, settled separately or ex post.

These considerations suggest the three-phase approach outlined below.

3.1 Phase 1 – A real-time market with a financial forward reserve market

This phase will simulate real-time dispatch and settlement rules for a sequence of periods and multiple locations. Physical characteristics including unit limits, ramp rates, and others that

separate spinning and offline reserves will be modeled. Parameters applicable primarily to unit commitment will not be modeled. Forward reserve positions of participants will be treated as parameters that differ among different scenarios. To investigate long-run effects, forward supply of offline reserves might be assumed perfectly elastic when the total remuneration reaches the carrying cost of a CT.

First, the model will be benchmarked to actual data for the chosen periods. Then the operations of the real-time market will be simulated according to the proposed rules. The results will be compared with the benchmark in terms of system dispatch, spot prices for energy and reserves, and other measures. Finally, sensitivity tests will be performed for key issues.

A main issue is revenue adequacy for offline reserves. The introduction of demands for reserves and settlements that provide capacity adders may create a situation in which the energy price (PE) is less than the sum of the marginal cost (MC) and the capacity adder (CA) for offline reserves but is greater than the sum of the marginal cost and the commitment cost (CC = start-up + no load). In that case, the unit will remain offline (since PE < MC + CA), but the offline reserve may qualify for a "lost opportunity cost" payment (PE – MC – CC). The probability of receiving such a payment increases with CA. The test may reveal how the setting of CA can impact revenue adequacy and thus investment incentives for the ICU peakers (the chief candidates for offline reserves) and baseload units. In the long term, revenue adequacy impacts resource adequacy of reserves via participants' investment decisions.

3.2 Phase 2 – A real-time market preceded by a physical forward reserve market, with multiple settlements

In this phase of the study, the simulator will settle the forward reserve and real-time markets in sequence, and allow intervening contingencies. This will allow investigation of further motives for reserves derived from the role of contingencies, such as uncertain loads and forced outages of generation and transmission. An important element is that reserve prices would include the option values of unit flexibility or availability associated with reserves, especially offline reserves.

The following are some relevant questions that the simulation can address:

- What is the relationship between the locational prices of forward reserves and those of real-time energy?

- What is the implication of including operability obligations in the ICAP specification so that offline units that cannot start are liable for the spot shortage price, presumably the TMOR price, and other penalties?
- How would the system perform under alternative definitions of offline reserves and schemes for sequencing the multiple bidding, clearing, and settlement processes for the products traded?

3.3 Phase 3 – Inclusion of unit commitments, with single or multiple settlements

Unit commitment introduces an additional dispatch and an additional payment to the single- and/or two-settlement market. The additional payment is the unrecovered portion of start-up and no-load costs. There are also ramping, minimum run-time and minimum down-time constraints on the schedules of individual units. These additional features will be accounted for in the market optimization and clearing model, and the resulting schedules of those units that are not self-scheduled. This phase will analyze agents' possible gaming strategies that cause such parameters and constraints to vary. The simulator will be enhanced to enable agents to bid startup and no-load costs, and to clear and settle the day-ahead unit commitment.

- If there is a single ex post settlement of reserves but separate day-ahead and real-time settlements of energy, then the most important task will be to study day-ahead bidding strategies when agents must take account of the combined effect of real-time energy and reserve prices. This study can be conducted with or without virtual bidding in the day-ahead markets.
- The portion of an ICU's (usually scheduled only for offline reserve) total revenue obtained from the day-ahead market can be identified.
- In this phase, the effect of the proposal for scarcity pricing of locational reserves in import-constrained areas can be investigated, since it uses re-optimized unit commitments as a means of estimating a lower bound on the scarcity value of locational reserves.

3.4 Overall scope of the simulation results

From these simulation experiments, one should be able to address several broad issues:

- Examine resource adequacy, operating, and reliability implications of the proposed market design.
- Obtain measures of the efficiency of the design, depending on various specifications (multiple settlements, financial or physical forward market for offline reserves, allowance for virtual bids in the day-ahead market, etc.). Rough predictions of the implications for energy and reserve prices can be obtained, and in particular, the relative effect of scarcity pricing of reserves can be identified.
- Identify the relationship between the existing ICAP market and the reserve markets in the proposed design. To what extent are they substitutes or complements in the reserve and energy markets? Obtain predictions about the effect on ICAP prices of reserve prices in forward and spot markets.
- Describe the main features of participants' optimal bidding strategies implied by the new design, and reveal where the design might be vulnerable to strategic behavior and gaming. Does the new design improve or impair the competitiveness of the energy and reserve markets?
- Predict the implications of the new design for adequate total revenues to owners of ICUs, and thereby provide indications about the design's prospects for ensuring adequate investments in offline reserve resources. This total revenue can be disaggregated into its constituent parts obtained in the sequence of forward, day-ahead, and real-time markets. Similarly, the total value of reserve resources can be disaggregated according to its constituent parts: covering large contingencies, quick response to deviations from forecasted system conditions, and risk management under uncertain system conditions.

Appendix 3: Settlements in a co-optimized market for energy and reserves

ISO-NE plans to introduce a day-ahead market in which energy generation and reserve capacity are co-optimized. Our understanding of the substance of the co-optimization is that:

- The optimization takes account of both energy and reserve constraints. Thus, scheduled generation must equal scheduled load, and also sufficient unloaded capacity must be available for each reserve category. A full network model is used for transmission. The reserve requirements include the first contingency and locational constraints. Virtual bids are allowed for energy (and might be allowed for reserves).
- The software uses integer programming (either mixed-integer or Lagrangian relaxation) to optimize the commitment of those units that are not self-scheduled. Also included are ramp constraints, EcoMin, and other technical data. For each unit, as-bid cost data include startup and no-load costs as well as a single supply curve (representing marginal costs of generation) that applies to every hour of an entire day.
- Transactions are settled by using the shadow prices on constraints. Since every bus in the system is represented, the shadow prices implement nodal pricing. However, in addition, if the daily cost of a scheduled unit is not fully recovered by the nodal prices then the deficiency is reimbursed and uplifted.
- After the day-ahead market, unscheduled units can re-bid in a market for residual unit commitments that are undertaken to ensure enough spinning reserve to cover the difference between forecasted and scheduled load (typically 10% currently), and reserves for the second contingency. Subsequent scheduling by the RAA copes with changes in conditions and forecasts since the DA market.

In this section, we focus on the implications of co-optimization for adequacy of offline reserve resources, especially fast-response peakers such as ICUs. In the New England system, the first contingency imposes stringent requirements for reserves. Half is allocated to spin and half to non-spin; also, half of the second contingency is met with 30-minute operating reserve and a quarter with 30-minute replacement reserve. The extremely large first contingency in New England (nearly 10% of peak load) is unusual. Its magnitude implies that, from a reserve

viewpoint, the New England system is ramp constrained, with tight requirements for fast-response reserves from online spin (including hydro) and offline hydro and ICUs.

Historically the problem in New England has been that the remuneration to ICUs scheduled for offline reserves has been insufficient to recover their long-run average costs. Recognizing this, ISO-NE recently introduced a forward market for offline reserves. We argue below that co-optimization in the DA market will not necessarily eliminate this problem unless settlements recognize in some fashion the value of ICUs in reducing the cost of unit commitments.

3.1 A simple example

Suppose first that the unit commitments are fixed. Then in the co-optimization, for each unit the capacity (between EcoMin and EcoMax) is allocated between energy generation and the various reserve categories. For simplicity consider only a single reserve category, which is spin (which can substitute for non-spin) for an online unit or non-spin for an offline unit. Also, ignore locational requirements and the locational effects of nodal pricing.

In a given hour, let CE and CR be the marginal cost of a specified unit for a given 1 MW of energy and reserve, respectively. Let PE and PR be the shadow prices on the energy and reserve constraints, derived from a linear-programming formulation of the co-optimization. Then the available 1 MW of capacity is allocated to either energy generation or reserve duty (but not both), depending which of the quantities (PE – CE) and (PR – CR) is larger, or that 1 MW is unused if the maximum of these two quantities is negative. For example, successive MWs of an online unit are typically scheduled for energy generation up to the point that PE = CE, then additional MWs may be allocated to spin (or nonspin) up to the point that PR = CR, and then residual MWs are unscheduled. An offline unit is scheduled for non-spin up to the point that PR = CR, and residual capacity is unscheduled. The settlement provides a payment of PE × QE + PR × QR, where QE and QR are the number of MWs scheduled for energy and reserve.

The key point of such a simple example is that the New England market provides no opportunity for any unit to bid a marginal cost CR of reserve availability. In fact, CR = 0 is always assumed, consistent with its true real-time cost. Therefore, offline units receive a positive payment (PR > 0) for reserve availability only when offline capacity is insufficient to meet the reserve requirement. In typical operations the reserve price is PR = 0. Therefore the payment to offline ICUs is typically zero when offline capacity is sufficient to meet the reserve requirement,

and an ICU must recover its long-run average cost almost entirely from sales in the real-time market when it is called in times of scarcity of energy supplies.

3.2 A more accurate example

In fact, the co-optimization also includes unit commitments. Each unit committed brings with it a reduction in the energy constraint by the amount of its EcoMin (and thus tends to lower the energy price PE). It also provides an amount (EcoMax – EcoMin) of available unloaded capacity that can be scheduled for energy, spin, or nonspin, and thus it can effectively substitute for any offline unit in the provision of non-spin and operating reserve capacity. When the available offline capacity is scarce, and the reserve price would be positive (PR > 0) the co-optimization will often commit an additional unit, and thereby drive the reserve price to zero (PR = 0). The price PR is depressed because the large chunk (EcoMax – EcoMin) of unloaded capacity from the committed unit removes the scarcity of offline capacity (since the committed unit's online dispatchable capacity can be allocated to any reserve category, including all those for which offline units can be scheduled). Thus, a major effect of unit commitments is that they reduce the prices of both energy and reserves. On the other hand, each unit commitment entails costs of startup and no-load, unrecovered portions of which are subsidized and the expense is uplifted.

An alternative view of this scenario is that ample offline capacity obviates the need for some unit commitments, and thereby possibly reduces the uplift of unrecovered startup costs. In an ideal settlement process, this savings in uplifted subsidies would be recognized and included in the remuneration paid to offline units. To some extent, this is the purpose of the forward market for offline reserve capacity that ISO-NE introduced recently.

Within the formulation of the co-optimization, one way to recognize the full costs of unit commitments is to include terms that explicitly recognize their effects. Thus, the objective function includes a term TC(NU, QU) for the total cost of providing a number NU of unit commitments, and a total quantity QU = NU × (EcoMax – EcoMin) of unloaded capacity from unit commitments (all committed units are assumed here to all be similar in cost and dispatchable quantity). Similarly, the energy requirement is reduced by NU × EcoMin, and the reserve constraint allows the quantity QU of unloaded capacity to be assigned to either energy, or to reserve duty within the limits that can be ramped in the required time of 10 minutes or 30

minutes for non-spin and operating reserve. At this writing, we are unsure whether this accurately depicts the intended formulation of the co-optimization that is planned by ISO-NE, but it will serve to illustrate the main points below.

3.3 Implications for settlements

The chief implication of the revised formulation above is that, in principle, the shadow price PR on the reserve constraint must satisfy the optimality condition that PR = MC(NU,QU), where MC is the marginal cost of additional MWs of the unloaded capacity QU of committed units. This "in principle" condition cannot be operationalized, however. The total cost function TC is actually a step function: it is discontinuous at each step corresponding to an additional unit commitment, and flat in the interval of length (EcoMax – EcoMin) between each step. The marginal cost is zero within an interval, and at each step, the marginal cost is actually a jump in the total cost by the amount of an online unit's startup costs.

If there were many small unit commitments, then one could approximate the total cost function, TC, continuously and thereby obtain an average measure of the marginal cost, and thereby an average measure of the value-added by offline capacity from ICUs that can substitute for unit commitments. If the co-optimization uses Lagrangian relaxation then, in effect, a similar kind of approximation is obtained. If mixed-integer programming is used, however, then no valid approximation is obtained, and typically the reserve price PR does not reflect the savings in the cost of unit commitments obtained from offline resources like ICUs. Also, the savings in uplift from offline resources is not otherwise credited to them in any other part of the settlement process.

Efficient management of the New England system entails two aspects. One is the efficient mix of online and offline resources. This mix is important both to provide spin (especially to cope with under-scheduling of load) and to provide sufficient reserve resources, either online or offline, to meet the large first and second contingencies. The second is a settlement rule that attracts investments in capacity to provide both the required magnitude and the efficient mix of resources. The distortion of DA reserve prices caused by unit commitments presents a basic dilemma. If no measures are taken to remunerate offline resources adequately, then investments in units like ICUs will be insufficient, and because they are insufficient, uplifted costs of unit commitments will be excessive—as in fact was the situation before the introduction recently of

the forward market for offline reserves. The forward reserve market provides one possible solution. However, if that market is essentially financial (as opposed to the physical and un-tradable reserve obligations now used), the ability of other units to compete in that market likely implies that the forward price will be no higher than the expectation of the DA price of offline reserves. The forward price thus would remain inaccurately low compared to the implicit savings in unit commitment costs. In principle, a financial forward market is preferable because it is more likely to attract an efficient mix of reserve resources (in the current forward market, the ISO specifies directly the amount of offline reserves it wants to procure via semiannual contracts). A suggested alternative is to internalize the cost of unit commitments (now subsidized via uplift) by declaring that units selling reserves in the forward market are ineligible for reimbursement of commitment costs.

3.4 Connections between the DA and forward markets for offline reserves

Co-optimization is intended to ensure an optimal mix of scheduled online and offline resources to meet both energy and reserve requirements in the DA market. However, this optimal mix is obtained each day from the existing portfolio of different kinds of units participating in the DA market. Over a longer time horizon, co-optimization in the DA market can ensure investments that result in an optimal portfolio of units of the various kinds only if the associated settlement process provides remuneration that accurately reflects the value-added, or cost-reduction, provided by each kind of unit. There is nothing inherently deficient in the daily minimization of the costs of meeting energy and reserve requirements obtained from a co-optimization process in the DA market that uses integer programming techniques. The focus must instead be on the settlement process, and in particular, on the inability of integer programming to provide a fully accurate measure of the marginal or average value of offline capacity that substitutes for unit commitments in large discrete amounts. Lagrangian relaxation methods of integer programming provide an imperfect measure of this value added, but mixed-integer programming typically provides no measure above zero.

The forward market for offline reserves provides a direct mechanism for ICUs to capture their value added if the reserve obligation is physical and un-tradable. If the ISO specifies the right amount of offline reserve procurements then the quantity and mix of resource types is de facto optimal. A prominent criticism, however, is that the optimal quantities and mix of resource

types can be assured only if shoulder units can compete with ICUs in the forward market. This seems to depend on making the forward market financial, in the sense that reserve obligations can be traded, or a provider can purchase replacement reserves in the DA market. However, if the forward market is financial, then arbitrage is likely to ensure that the forward price is no more than the expected DA price of offline reserves, which would remain below the actual value added of offline capacity. It seems evident that a closer approximation of value added will be obtained only if unit commitment costs are internalized—since the subsidy of un-recovered startup costs is the ultimate source of the under-valuation of offline reserve capacity. The proposal to require that units providing long-term offline reserves must be self-scheduled, and therefore ineligible for reimbursement of un-recovered startup costs, provides one way to internalize a portion of the unit commitment costs. However, this proposal internalizes only the commitment cost of a unit participating in the forward market—it does not reflect the subsidization of other units that remain available for unit commitment and scheduling in the DA market and that therefore continue to depress the DA price of reserves (and energy).

A remaining possibility is that settlements in the co-optimized DA market reflect more than just the marginal shadow prices on constraints. For example, the co-optimization might produce as an output a measure of the *average* (as opposed to the *marginal*) value of offline capacity as a substitute for the uplifted cost of unit commitments that would be required if the offline capacity were absent. If an appropriate measure of average value added were used then the remuneration paid to offline units would presumably be closer to the ideal remuneration that would, over the long term, attract the optimal quantities and mix of resources competing to fill the ISO's daily requirements for offline reserves.

Appendix 4: Including out-of-merit commitment costs in the reserve price

Here we examine a procedure for calculating a price for local reserves as outlined in the memorandum by John Farr, "Using Out-of-Merit Energy Costs as the Basis for Setting Reserve Prices," 2/23/04. We focus mainly on the penultimate section on "A Comprehensive Methodology."

4.1 Purpose of the procedure

The purpose of the procedure is to establish a price for TMOR in real-time that can be used in settlements for all units assigned to that reserve status in an import-constrained local area.[14] This purpose presumably reflects the expressed intention of ISO-NE to provide reserve capacity with compensation that reflects scarcity to some degree, on the hypothesis that payment of some portion of scarcity rents might favorably affect investments in new capacity in import-constrained local areas. (New England is unlike some markets elsewhere, such as California, that use bid-based day-ahead markets for reserve procurements that enable generators to capture some portion of scarcity rents.) Farr recommends a procedure in which the average costs of TMOR ($/MW/h) of those units that either were, *or might have been,* committed in a particular hour are used to find a *lower bound* on what the actual costs might have been, and hence a lower bound on what the scarcity rent might be. The procedure is applied ex post; that is, whatever the actual dispatch was in a given hour, after the fact the available bids are re-analyzed to determine an appropriate price of reserves that in settlements is paid to all actual reserve capacity in that hour.

4.2 Main ingredients of the procedure

Farr's prevailing assumption is that all designated TMOR reserve capacity in a given hour will be paid a uniform price, measured in $/MW/h. This price is paid only to those units assigned to TMOR status in that hour (in addition, each actually committed unit is assured full recovery of its as-bid costs). Since the reserve price is calculated ex post, the energy price used in the calculation is the one that actually prevailed in that hour. The units actually assigned to TMOR status in that hour need not be the ones that would have been optimal in that hour based on the

[14] A theoretical justification for positive reserve payments when reserves are scarce is developed by Paul Joskow and Jean Tirole, "Reliability and Competitive Electricity Markets," MIT, March, 2004. Besides a scarcity value of reserve capacity, the formula they derive includes an additional term whenever there is risk of grid collapse.

data used for the calculation of the reserve price. The motive for this restriction is the concern that units may actually have been committed due to intertemporal constraints (such as minimum run times for some units) or other considerations; therefore, an ex post calculation based on the data for the given hour need not imply that the actual dispatch was optimal based solely on the data for that hour.

4.3 The basic procedure for real-time ex post settlements

The basic procedure is the following. Settlement software receives:

- The actual energy and reserve requirements for the designated hour in the local area.
- The actual dispatch and the actual energy price for that hour in the local area.
- For each unit in the local area, some technical constraints (EcoMin, EcoMax, ramp rate, quick start capacity), but not others (e.g., minimum-run-time is not used in calculating settlements).
- The bids for that hour from all units in the local area, whether actually dispatched or not. These bids include:
    - No-load cost
    - Energy cost, as bid, for each block of capacity

From this data the software calculates:

- The energy and reserve requirements actually needed to be filled in the local area, net of actual imports and unused tie capacity.
- For each unit the quantity (MW) of TMOR that it was capable of providing. This includes two components, each calculated net of its actual energy dispatch (DDP):
    - Its quick-start capacity and/or that portion of its capacity that it can ramp in 30 minutes without exceeding EcoMax.
    - The energy provided by its EcoMin that would enable other units to be backed down to provide TMOR (actually, spin).
- For each unit the total cost ($/h) of its TMOR capability, obtained as the sum of its no-load cost (if absence of quick start capability, and a positive EcoMin, requires that it be committed) and that portion of its energy cost that exceeds the actual energy price in the local area (i.e., out-of-merit energy cost).

- For each unit the ratio ($/MW/h) of the previous two items, that is, its average cost of TMOR capability.

This data is analyzed as follows to obtain a price:

- The units' blocks of TMOR capacity are arranged in the merit order of ascending average costs of TMOR capability.
- The price is chosen to be the average cost of TMOR capability for the marginal block of capacity were the blocks dispatched according to this merit order until the net requirement for local TMOR supplies is met or minimally exceeded.[15]

Finally, the settlement pays this price to each MW of capacity actually designated for TMOR status in the local area in the given hour.

4.4 Main properties of the procedure

Farr makes two claims about the features of this procedure. In our words, and with our comments appended, they are:

1. It is appropriate to allow that the actual dispatch and dispatch implied by the merit order can differ.
    a. Our comment: Given that a procedure is to be applied ex post for settlement in each hour separately, this feature is inevitable. In contrast, if a similar procedure were applied to day-ahead settlements then one would want the day-ahead dispatch and merit order to be more closely aligned – and ideally one would use the shadow price on the local reserve constraint as the price, perhaps obtained via a Lagrangian relaxation algorithm.
2. The price calculated via the above procedure provides a "defensible" lower bound on the average $/MW/h cost of TMOR capacity from the last block in the merit order.[16] This claim derives from two considerations:

[15] In connection with Example 2 (page 2), Farr considers an alternative calculation of average cost that uses assigned TMOR capacity in the denominator. Similarly, in connection with Example 1 he proposes that only actually assigned TMOR capacity be paid the price, not the available TMOR capacity of an assigned unit. He also suggests a payment rule (page 4) that penalizes a unit from deviating from its DPP.

[16] Farr notes correctly that no-load costs can have the effect that blocks of the same unit will not be ordered monotonically. We ignore this consideration here since in practice one expects that an integer-programming algorithm would be used to implement the procedure, thus ensuring a global optimum for which such considerations are moot. Farr's reliance on a monotonic merit order should be interpreted as a plausible simplification for the purposes of exposition of the main ideas. The integer-programming algorithm would presumably be a reduced form of the one actually used for dispatch, with deletion of intertemporal constraints and with a fixed energy price.

(i) Intertemporal constraints and other considerations affecting the actual dispatch would imply higher scarcity rents then those obtained from an optimal dispatch based only on the data for the given hour.

(ii) Using a merit order constructed from all available capacity, not just the capacity actually dispatched, ensures that the possibility of substitution of a cheaper source of TMOR is considered. Thus (as in Farr's Example 5) a unit that was not actually dispatched might have been cheaper (on an average cost per MW basis, ignoring intertemporal constraints) than one that was actually dispatched – in which case the price obtained by the above procedure is generally less than one would obtain by considering only the blocks actually assigned to TMOR status.

a. Our comment: This claim is valid to the extent one accepts average cost on a $/MW/h basis as a valid approximation of scarcity rents. As Farr points out, lumpy capacity (due either to EcoMin or no-load cost) can result in large average cost for a small quantity of TMOR, in which case some upper bound is required, presumably no more than the price cap.

We see these two features as the distinguishing elements of Farr's proposal. The first proposes an ex post re-optimization of TMOR assignments within each hour in each local area as a way of estimating a lower bound. The second proposes that using a simple merit order based on the average cost of each block available for TMOR will provide a lower bound on scarcity rents, absent intertemporal constraints. As mentioned in footnote 2, the merit order based on average costs should be considered an expositional device, since an implementation would presumably use integer programming. There are other practical considerations too, since an average cost can be very large if the denominator (available or required TMOR capacity from the unit) is small.

Our summary view of the proposal is that, in the context of ex post settlements:

- An ex post re-optimized within-hour assignment of TMOR duties is potentially a valid way to obtain a lower bound on the costs of meeting TMOR requirements within each hour.
- A merit order based on each block's average cost of available or assigned TMOR capacity would not actually be valid in practice if no-load costs and EcoMin constraints are significant, but this is easily remedied by using an integer programming algorithm for settlements that is adapted from the one actually used for dispatch.

4.5 Summary Remarks

Farr's proposal assumes implicitly that the purpose of such a procedure is to obtain a per-MW price for actually dispatched TMOR capacity that reflects to some degree the scarcity rent on available capacity. Due to units' technical constraints, no mathematical theory suggests that a uniform per-MW price is a valid measure of scarcity rents (even absent intertemporal constraints), but we acknowledge that there are other reasons for using this form of capacity payments for TMOR.[17] Given this proviso, the residual issue is whether a lower bound of the sort proposed by Farr is a sufficient payment to attract efficient levels of investment in reserve capacity in local areas. One can be hopeful, since usually the price produced by the procedure would be lower than the actual average cost of a marginal block, if at all, only due to intertemporal constraints. However, only experience can reveal whether a "signal" of this sort is effective. There are two main reasons for pessimism. One is that this price is not assured to be sufficiently large sufficiently frequently to attract resources of the kinds most needed by the ISO (ICUs being the prime example). The second is that scarcity rents in a local area might disappear immediately upon completion of a new unit within the area, and anticipating this, an investor would not consider revenue from the TMOR capacity price to be a relevant consideration in deciding on whether to install a new unit.

[17] The closest approximation are Vickrey prices, but they are not uniform; that is, each unit receives a compensation that is equal to its *total* contribution to reducing the ISO's procurement costs. This total amount need not be computable as the product of its capacity assigned to TMOR and a uniform price.