NEW ENGLAND DEMAND RESPONSE INITIATIVE

LONG-TERM RESOURCE ADEQUACY: THE ROLE OF DEMAND RESOURCES*

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

November 3, 2002

DRAFT

*This draft paper was prepared under contract with Lawrence Berkeley National Laboratory, Berkeley, CA (Charles Goldman, project manager) for the New England Demand Response Initiative.

1. INTRODUCTION

Can energy markets alone provide sufficient information and incentive for the construction of enough new resources to meet society's needs for reliable and reasonably priced electricity? To some industry observers, the California electricity crisis of 2000 and 2001 suggests that additional measures are needed to maintain reliability (especially to prevent blackouts), prevent the undue exercise of market power by the owners of generating units, and ensure that electricity prices do not rise to levels that are unreasonable (California Energy Commission 2002). These additional measures involve a long-term commitment to resource adequacy to ensure that, in real time, enough generation and demand-response resources will be available to meet the needs for energy, congestion management, and ancillary services.

It is essential to distinguish between *planning* reserves and *contingency* reserves. The first relates to long-term adequacy and the second to short-term security.[*] For day-ahead planning and real-time operation, system operators are required by the North American Electric Reliability Council (NERC) and regional-reliability-council rules to maintain minimum levels of contingency reserves, usually 5 to 8% of the projected daily peak. (The minimum contingency-reserve requirement is typically based on the size of the largest generating unit or transmission element online within the control area.) These short-term reserves, which must respond fully within 10 to 30 minutes of being called upon, protect bulk-power systems from the effects of major generation and transmission outages and correct for errors in day-ahead load forecasts.

Planning reserves (of which contingency reserves are a subset) provide long-term insurance against problems that might otherwise arise when units are not available (e.g., for planned maintenance) and allow for unanticipated long-term load growth. Generator outage rates are on the order of 5 to 30%; that is, units are available 70 to 95% of the time. Planning reserves provide sufficient capacity to offset these planned and sudden losses.

The further into the future one is planning and assessing reserves, the greater the uncertainty about the real-time availability and operation of these resources (Fig. 1). For example, three years ahead of the operating day, the resource may be a plan, not a physical unit

[*]The definition of reliability used by NERC (2002) encompasses two concepts, *adequacy* and *security*. Adequacy is defined as "the ability of the system to supply the aggregate electric power and energy requirements of the consumers at all times." NERC defines security as "the ability of the system to withstand sudden disturbances." In plain language, adequacy implies that there are sufficient generation and transmission resources available to meet projected needs plus reserves for contingencies. Security implies that the system will remain intact even after outages or other equipment failures occur.

on the ground. Such a planned resource is subject to many uncertainties associated with project financing, siting and permitting (related to environmental and land-use issues), the delivery of equipment from manufacturers, construction labor, interconnection to the transmission network, and so on. Six months ahead of the operating day, the uncertainties are greatly reduced; at this time, they relate primarily to planned-maintenance and forced outages, input fuel prices, and electricity prices. During and just ahead of the operating day the only uncertainty is whether the unit will suffer an equipment failure leading to a forced outage. These uncertainties, unlike those associated with load forecasts (discussed later) are asymmetric. That is, the uncertainty is almost always that the resource will provide less, not more, than committed ahead of time.

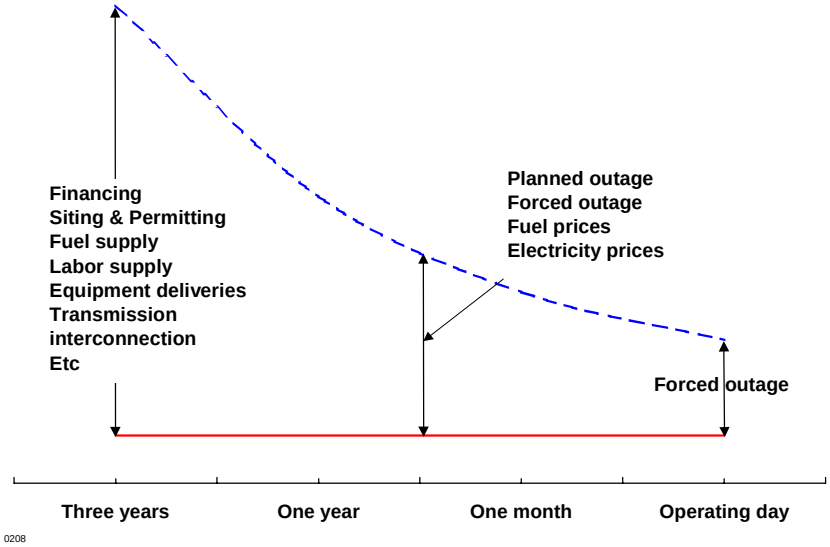


Fig. 2. Schematic showing how uncertainty about real-time availability and performance increases with the amount of time ahead of the operating day (real time).

This paper discusses long-term resource adequacy, beginning with a discussion of the underlying concepts. The paper then describes the proposal from the U.S. Federal Energy Regulatory Commission (FERC 2002b) in its July 2002 notice of proposed rulemaking on Standard Market Design (SMD). Section 4 investigates the different ways that various kinds of demand resources might qualify for and participate in markets for such resources. Section 5 reviews the programs run by the independent system operators (ISOs) in the northeastern United States and a proposal from the California ISO. The final section summarizes the paper and offers conclusions and recommendations.

2. UNDERLYING CONCEPTS

If we are serious about creating competitive wholesale electricity markets, why are we considering a regulatory approach that mandates minimum amounts of resources on a long-term basis? Why can't the spot prices produced by competitive markets (i.e., the day-ahead and real-time markets operated by today's ISOs and proposed by FERC in SMD) stimulate the appropriate amounts, types, and timing of new supply and demand resources?

In principle, these spot markets for energy, congestion management, and ancillary services (regulation and contingency reserves) should provide sufficient motivation for investors to build new power plants and transmission lines and install load-management and energy-efficiency equipment on the premises of retail customers Besser, Farr, and Tierney 2002; Shanker 2002).* In such an ideal world, relative to one with mandated long-term reserve margins, electricity prices and costs would be lower, energy prices would be more volatile, and the mix of generation would include more baseload units and fewer peaking units (Hirst and Hadley 1999).

For spot markets to perform this essential investment-motivation function, however, the markets must be competitive. Such competition implies many buyers and sellers, none of which is able to influence market prices, demand and supply that respond to prices, no barriers to entry, and functional market designs. Unfortunately, the U.S. wholesale markets operating today do not fully meet these requirements. Even the New York and PJM designs, generally acknowledged to be the best in North America, lack all the features of fully competitive markets.

In particular, almost all retail demand neither sees nor responds to time-varying wholesale prices. Specifically, most customers have meters that record monthly, not hourly, electricity consumption and, therefore, are unable to respond to changes in prices from hour to hour. This demand inelasticity prevents markets from working efficiently when scarcity occurs.

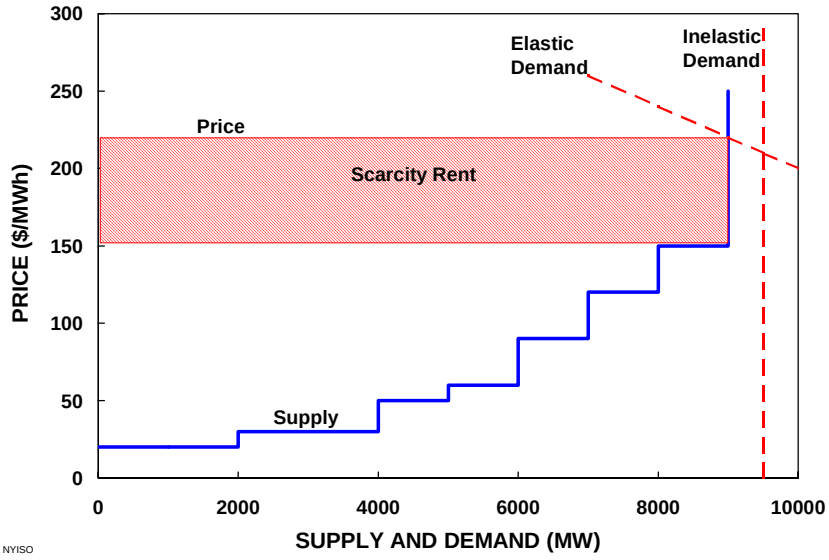


Fig. 3. **When demand is completely inelastic and exceeds available supply, no market-clearing price exists. However, when demand is elastic, it sets the price during shortage conditions. The difference between the price of the most expensive unit online and the demand curve is considered the scarcity rent.**

Without any demand response at all, electricity markets would find no market-clearing prices when shortages occur (Fig. 2). Thus, price caps are

*For example, between July 1, 2001 and June 30, 2002, the real-time energy price in PJM exceeded $100/MWh 1% of the time, with an average price during those 92 hours of $320/MWh. This price is far above the variable cost of even the most expensive generating units.

essentially a requirement in spot electricity markets with no demand response. With some demand elasticity, however, the price can rise above the bid of the most expensive generator then online. This extra payment, shown as the shaded region in Fig. 2, is paid to all generators, which should motivate construction of new resources.

In addition to a general lack of demand elasticity, existing markets contain other flaws that prevent prices from reflecting fully the value to consumers (and, therefore, the incentive to investors to build more infrastructure). Because of deficiencies and market rules, the existing ISOs all have bid and price caps, approved by FERC, that prevent prices from reaching unreasonably high levels. Unfortunately, spot prices must occasionally reach such high levels if these prices are to stimulate needed investment.

Also, the rules used to set prices sometimes do not reflect the physical reality of current grid conditions. Consider, for example, a situation when generation is lagging demand and the only resource that can provide enough capacity quickly enough is a block-loaded combustion turbine (CT). Such a unit may have a running cost much higher than that of the other units then online. But, because the unit is block-loaded, the market rules in effect may not consider the unit to be marginal, making it ineligible to set the market-clearing price. So, a cheap unit that is backed down to accommodate the CT's capacity sets the price. And the startup and operating costs of the CT are recovered through an uplift charge. In other words, the need for additional capacity requires the ISO to start and operate an expensive generator, which in turn suppresses the spot price of energy.

Both the New York ISO and ISO New England faced situations such as this. As Patton (2002) noted, "The analysis of these peak periods indicates that energy prices were established at inefficiently low levels. … [T]he underlying cause of these price effects can generally be traced to the current [ISO-NE] market rules and procedures." Both ISOs modified their market rules to ensure that the CT can set the price when its capacity is needed. ISO New England (2002a) proposed "[p]ermitting peaking units and external contracts that can be committed and decommitted hourly (though not dispatched [during the hour]) to be considered in the calculation of the Energy Clearing Price (ECP) for those five-minute intervals when they are economic for energy or required to supply spinning reserves or Operating Reserves in shortage conditions;" FERC (2002a) approved these changes.

These deficiencies in current market designs and operation, coupled with generator market power, price and bid caps, insufficient price-responsive demand, and other problems suggest that some long-term mechanism is needed to ensure the existence of enough supply and demand resources to prevent shortages. Unfortunately, designing a long-term product to meet this need is far from simple.

The primary question concerning resource adequacy is "What is the product?" What characteristics must the resource exhibit over different time frames, from real time operations during an emergency to day ahead and, perhaps, months and years ahead? In real time, the

resource must be physical and fully available. Three years ahead of operations, however, the resource might be considered qualified if it appears in a credible plan for construction or acquisition.

Is the resource-adequacy product a form of physical insurance or financial insurance? System operators (and, perhaps, government regulators) want assurances that, when needed, enough resources will be available and will respond to operator requests for energy or reserves to maintain reliability, a form of physical insurance.[*] But such insurance can never be available 100% of the time. As the time grows between declaration of a resource and system-operator dispatch of the resource the uncertainty about the availability of the resource also grows (Fig. 1). For example, generator owners bidding into an ISO's day-ahead markets for contingency reserves know that their units will not be on a planned or maintenance outage and that they have sufficient fuel and labor resources to operate the facility. Although a forced outage might occur between the time the day-ahead market closes and real-time, this is a low-probability event. At the other end of the temporal spectrum, a declaration of capacity, either supply or demand, three years ahead of operation is highly uncertain and subject to problems related to fuel supply, equipment failures, economics, labor problems, and so on. Thus, the amount of capacity to be acquired for a particular event (or time) is larger the further ahead of time the resources are reserved. For example, a real-time contingency-reserve requirement equal to 6% of that day's peak load might translate into a long-term planning-reserve requirement of 15% of expected summer peak load.

It is also unclear whether long-term resource adequacy is needed to deal with reliability or economic conditions. Is resource adequacy:

- a solution to a reliability problem (i.e., the need to have enough resources in real time to meet the needs for energy, losses, congestion management, regulation, and contingency reserves)?

- a solution to a market problem (i.e., the need to have enough resources participating in day-ahead and real-time markets to prevent the exercise of market power)?

- needed to correct deficiencies in current market design (e.g., price caps that prevent spot prices from reaching their market-clearing levels during periods of shortage and market rules that suppress prices during shortage periods)?

Although these three objectives differ enough that the preferred solutions could also differ, they all involve the availability of enough physical resources (generation and enough transmission

[*]When a homeowner buys fire insurance, the insurance company does not guarantee that the house will not burn down (which would be a form of physical insurance that could be provided by stationing fire fighters in the home all the time). The insurance company guarantees to pay the homeowner for the damage associated with loss of the house and the household belongings, a form of financial insurance.

to deliver the output and/or demand) to meet customer demand and secure the bulk-power system in real time. In some cases, this may require investment in energy infrastructure as well as the maintenance of equipment that might otherwise be retired as uneconomic.

In general, long-term resource adequacy is a commitment on the part of the resource owner to convert the committed capacity into energy (or load reduction) upon demand. In the short-run (e.g., day ahead), physical resources are needed to make this commitment meaningful. In the long run (e.g., three years ahead), however, it may be sufficient to have a credible plan to build or otherwise acquire such a resource. Complications arise because of the uncertainties over converting plans into reality.[*] As NERC (2002) noted:

> The announcement of a new merchant generating facility does not necessarily guarantee its construction for a variety of reasons, including future market prices, the ability to obtain suitable interconnection and transmission access agreements, and the ability to obtain financial backing and other business-related factors. In some cases, a single developer may announce several alternative projects, even though only one will be built. Such announcements are made because developers cannot be assured of obtaining all the necessary permits to build a power plant at one location, forcing them to consider alternate locations as a contingency plan. In other cases, economic or political conditions may change, making a project unprofitable, leading to its cancellation. For example, volatility in natural gas prices may cause developers to review previously announced plans to construct new gas-fired generating units. Similarly, the institution of price caps for wholesale electricity sales also may lead to project cancellations. Finally, some states have issued moratoriums on new power plant construction because the capacity of the proposed facilities exceeds the projected future demand for electricity in the state or out of concern about the environmental consequences of hosting generating facilities whose output could be sold out of state.

In addition to questions about product definition and purpose, one can ask what other industries sell the rights to their output so far ahead of actual product delivery. We pay for restaurant meals after the food is delivered and eaten, we pay for a visit to the doctor after the consultation is complete, and we pay for hotel rooms as we leave the hotel. We pay for movies, car insurance, and airline tickets ahead of time, but usually immediately before taking delivery of the product or service. I often buy airline tickets a few weeks ahead of time. But in that case, I am buying a well-defined product (e.g., Seat 26C on Delta's flight 639 from Cincinnati to Denver on October 24, 2002). I can think of no other product that is purchased so far ahead of time with so little clarity on the product being purchased and its delivery date. This raises the

[*] For long time periods (e.g., two years) should the resource-adequacy requirements focus more on plans and reporting and less on actual demonstration of resource capability? At what point (e.g., month ahead or day head) should the emphasis shift from plans to physical reality?

question of what is so unique about electricity that it requires a separate market for the capability to provide the services consumers want rather than rely solely on the markets for those services.[*] And, why does the system operator (or government regulator) *require* load-serving entities (LSEs) to purchase this product?

Another critical question is who determines, and on what basis, how much capacity is enough. Clearly, requiring higher reserve margins improves reliability. Just as clearly, higher reserve margins increase electricity costs to consumers. Stoft (2002) notes that the commonly used criterion for loss of load probability, one-day-in-ten-years, appears to have no technical or economic basis.[#]

Should the resource obligation be based on the annual system peak, seasonal peaks, or monthly peaks? In particular, would obligations based on the projected annual peak (e.g., summer) impose costs on consumers, without additional reliability benefits, during nonpeak months (e.g., winter)?

In addition, load forecasts, a key factor determining each LSE's resource obligation, will almost surely turn out to be wrong three years later when the actual load materializes. Economic growth will be either higher or lower than forecast, the summer will be hotter or cooler than average, and other factors will have changed relative to the forecast. It is important to note that the load-forecast uncertainty is symmetrical; the forecast is as likely to be too high as it is to be too low. The resource uncertainty (Fig. 1), however, is asymmetrical. If the load forecast is independently prepared by the regional transmission organization (RTO), it is likely to be unbiased. However, if each LSE provides its own forecast (or the inputs the RTO uses to make a regional forecast), the forecasts are likely to be biased down. That is, LSEs have an incentive to underforecast to lower their long-term resource obligations.

The overall resource adequacy requirement is a function of two highly uncertain factors, the calculated reserve margin and the load forecast. To make matters worse, the result is the product of these two factors:

Resource Obligation = Forecast Load × (1 + Planning Reserve Margin).

---

[*]The inability to economically store electricity makes it different from most industries. However, airline seats cannot be stored either and we do not purchase the long-term rights to these seats.

[#]Determining the optimal level of installed capacity involves trading off the value of lost load against the cost of peaking units. Stoft (2002) states "This approach is never taken. Instead, the crucial input is the acceptable number of hours of load shedding, an engineering constant shrouded in secrecy but said to be 'one day in ten years.' [footnote] How a parameter describing consumer economics was derived from astronomical constants remains a mystery. Why should the cost-minimizing value of load shedding equal the time it takes the earth to rotate once times the number of digits on two hands divided by the time it takes the earth to orbit the sun?"

Although the above equation is quite simple, it is technically and economically difficult to evaluate the two components, forecast load and planning reserve margin. For example, setting the required reserve margin too high increases consumer electricity costs beyond what is necessary; setting the margin too low could threaten reliability. And the appropriate reserve margin depends on the extent to which retail demand responds to price; greater price responsiveness reduces the need for extra physical capacity.[*] The California Energy Commission (2002) used "a mixed approach that combines scenario analysis with probabilistic assessments" to determine how much installed capacity is enough. The Commission analyzed the effects of temperature on summer-peak demand and the effects of hydrological conditions, potential construction delays, and age-related forced-outage rates on supply adequacy.

3. FERC PROPOSAL

FERC (2002b) proposed "a resource-adequacy requirement to provide for sufficient supply [generation and the transmission necessary to deliver the generation output to load centers] and demand resources to avert shortages … [and] involuntary curtailments."[#] FERC stated that such a requirement is necessary because

> Most resources take years to develop and spot market prices [day ahead and real time] do not consistently signal the need for new infrastructure in the electric power industry. Moreover, spot market prices that are subject to mitigation measures may not produce an adequate level of infrastructure investment even after a shortage occurs. Further, as long as regional resources are made available to all regional load-serving entities and their customers during a shortage, such entities have the incentive to lower their supply costs by depending on the resource development investments of others, a strategy that leads to systematic under-investment in infrastructure by all load-serving entities in the region.

Thus, FERC recognizes that this deficiency in spot markets is a consequence, in part, of its imposition of bid and price caps in these markets. That is, if spot prices were permitted to rise to their scarcity value when shortages occur, potential investors might have sufficient incentive to build new power plants and transmission lines and to implement demand-response programs. However, customers (and, therefore, regulators and legislators) strongly dislike the very high spot prices that occur when unconstrained demand exceeds available supply.

---

[*] The California Energy Commission (2002) noted that "… there is as yet no reliability planning paradigm that accounts for the willingness of consumers to forego some electricity usage when prices are high, despite clear evidence that this is quite acceptable to many end users."

[#] The FERC notice discussed Long-Term Resource Adequacy in Section IV.J (paragraphs 457 through 550) and in Part IV.I of the proposed pro forma tariff.

In addition, the present-day inability of most retail load to see and respond to hourly changes in wholesale prices further complicates the maintenance of a suitable supply:demand balance. Finally, planning and building new energy infrastructure is a lengthy process, often taking several years from initial design to online operation. These factors could lead to boom-and-bust cycles in energy-infrastructure investments, leading to FERC's concern about "sustained periods of inadequate supplies, threatening the reliable operation of the bulk power system." A long-term resource-adequacy requirement could smooth out these cyclical patterns of overbuilding and underbuilding.

FERC believes that, under existing conditions, LSEs have an incentive to underinvest in new capacity or sign too few long-term contracts for the energy supplies needed to serve their loads. Rather, it may be cheaper for some LSEs to free ride on the resources that others acquired.[*]

FERC states that its proposal is like the traditional requirement to maintain a minimum planning reserve margin imposed by state regulators on utilities. It is, however, unlike the installed capability (ICAP) requirements currently in place for the three northeastern ISOs. FERC is explicit about its dislike of the current ICAP programs: "We are reluctant to impose a national ICAP requirement, in part because of our concern about the effectiveness of the existing ICAP programs and in part because they were based on the former voluntary tight power pools." However, FERC is largely silent about the particular aspects of ICAP that it dislikes. Conversations with FERC staff suggest that FERC is concerned that the price of installed capability in the ISO ICAP markets is almost always zero or very high, with few intermediate values, suggesting that there are no tangible benefits from ICAP; ICAP does not stimulate construction of *new* resources (i.e., it pays for both existing and new resources and its time horizon is only a few months out); and it is subject to market-power abuses. Thus, FERC's proposal is intended to "replace the current ICAP programs."

FERC proposes, instead, a longer-term process that might extend two to four years ahead of operations. While FERC proposes to allow each region to select an appropriate planning horizon, in all cases the time period chosen should be long enough to motivate and achieve construction of new generation and demand resources in time to avert shortages.

[*] In real time, it is virtually impossible for the system operator to interrupt service to those customers, and only those customers, for which sufficient capacity was not acquired before hand. Most customers are not connected directly to the bulk-power system and most transmission nodes are not amenable to remote switching. As a consequence, the system operator, when faced with the need to interrupt load, usually does so through the distribution utilities, which institute rotating blackouts among various feeders, each of which serves many retail customers.

The process begins with a long-term load forecast for the region, prepared by the RTO. The RTO then determines a suitable planning reserve margin, for which FERC proposes a minimum value of 12%,[*] with regions free to choose a higher level if appropriate.

The RTO's regional load forecast and minimum reserve margin then forms the basis for the amounts of resources each LSE would be required to acquire. Each LSE's obligation is to be based on either its recently documented load-ratio share or a forecast of its load-ratio share.

FERC is clear that both generation and transmission and demand response resources qualify for such long-term resources. FERC notes that "Better demand response to high prices when a shortage condition approaches will lower demand and reduce the use of high-cost power resources. Demand response will help ensure reliability, prevent a shortage that could produce a curtailment, act as a check against market power, and provide a yardstick for the value that buyers place on supply." Demand resources can be either "biddable and interruptible load," although FERC defines neither biddable nor interruptible.

The proposed standard for supply is that "the generation is physically feasible; that is, the generating units are capable of generating the power planned, and enough transmission is available to deliver the power from the generating station to the particular load." A contract with a power marketer, on the other hand, to deliver power from "unspecified sources cannot satisfy the requirement."

The standard for demand response states that these resources must be "verifiable," such that the RTO has "confidence that the demand response resource will be able to contribute when called upon during a shortage. Demand response may be obtained through biddable demand reduction, interruptible load, or other dependable load management program." The RTO might require "the load-serving entity to install equipment that gives it direct control over the loads of the customers that are subject to the interruption."

FERC proposes penalties to encourage LSEs to acquire the resources needed to meet their obligations. Curiously, FERC's penalties operate only in real time and only during emergencies. As FERC puts it, its penalties "occur at the end of the planning horizon, not at the beginning." The two-part penalty system would (1) curtail the loads of deficient LSEs first (where feasible) and (2) add a penalty rate to the real-time spot price, with the penalty rate related to the level of reserve shortfalls (e.g., a penalty of $500/MWh when contingency reserves are 1% below the required level and $700/MWh when contingency reserves are 3% below the required level.) Failure to comply with an RTO's interruption order during a shortage would result in a penalty of $1000/MWh (in addition to the real-time price of energy) for all energy taken by the offending LSE.

---

[*]This reserve margin is equal to 100×(installed capability/peak load − 1), where capability and load are both measured in MW. A 12% margin means that the amount of supply and demand capacity in place must exceed the projected peak demand by 12% or more.

Although these penalties ensure that deficient LSEs are penalized for the costs they impose on the system when shortages occur, the focus on real-time penalties is inconsistent with FERC's desire to encourage long-term investment. If the goal is to encourage investors to build power plants and transmission lines and to implement demand-response programs, the

Exhibit 1. Effectiveness of Real-Time Penalties

Are the real-time penalties proposed by FERC likely to be sufficient to motivate LSEs to invest in new generation and demand-response programs?

The annual carrying cost of a CT might be about $60,000/MW-year. How many hours of shortage would it take to exceed that level? Consider a combined spot price plus RT penalty of $1000/MWh. If the number of shortage hours is less than 60 (60 × $1000 = $60,000), the annual fixed cost of a CT), the smart LSE will face the penalties and not purchase long-term supplies.

As noted in the footnote on page 4, PJM experienced high prices for 92 hours between July 1, 2001 and June 30, 2002, with an average price of $320/MWh. If PJM faced shortages during all 92 hours (highly unlikely), the cost to a deficient LSE, exclusive of penalties, would be only half the carrying cost of a new CT ($29,400 vs $60,000/MW-year).

penalties should be levied far in advance of real time. That is, reliance solely on real-time penalties is likely to encourage some LSEs to free ride (exactly what FERC found problematic in the existing ICAP programs) on those LSEs that acquired enough resources. A risk-taking LSE might reason that enough LSEs will acquire enough resources to make real-time shortages very unlikely and short-lived occurrences. In such cases, it might be cheaper to pay the occasional real-time penalty and very high spot prices in exchange for avoiding the long-term costs of new capacity (Exhibit 1).

FERC permits but does not require the RTO to establish and operate markets for qualifying resources. PJM and New York currently run ICAP markets over various time frames, from months ahead, through day ahead. FERC appears to prefer bilateral arrangements among market participants without the RTO's intervention.

Finally, the FERC proposal is silent on whether and how the resource-adequacy requirements and resources specified, say, three years ahead of operation might be updated during the intervening three years. As discussed above, many factors will affect the extent to which plans become reality. Surely, LSEs should have the opportunity to revise and update their plans as events unfold.

4. ROLE OF DEMAND RESPONSE

Determining how demand resources can provide Long-Term Resource Adequacy is difficult for at least two reasons. First, FERC's explanation of resource adequacy is vague, as discussed in the preceding section. Second, FERC offers this new paradigm to replace the existing ICAP requirements, currently used by the three northeastern ISOs. FERC is clear that it sees problems with ICAP but its Notice does not explain what these problems are. It is difficult to determine how customer loads can provide such resources when the definition of these resources is not clear.

More fundamentally, demand reductions are different from generating units. Although the two sets of resources should be treated in a similar fashion, they have, in some respects, very different characteristics. Supply resources exist to provide energy; that is why they are constructed. Demand resources exist to produce goods and services or personal amenities (e.g., heating, lighting, and cooling); retail customers do not think of themselves as being in the energy-supply business. As a consequence, their willingness and ability to reduce load with little advance notice can be limited. Existing demand-response programs recognize these restrictions in that they generally require participating customers to agree to only a limited number of interruptions each season and for only a few hours at a time. Generators, on the other hand, are more than willing to produce energy hour after hour as long as the price is high enough to recover variable costs and contribute to fixed costs.

Perhaps the most important demand resource would be the creation of a demand curve for planning reserves. As noted in section 2, determining the "correct" amount of reserves to require is as much art as science. Clearly, if reserves are cheap, the ISO might want LSEs to carry more reserves than if they are expensive. Exhibit 2 describes how the New York ISO plans to implement this idea for contingency reserves.

The existing ICAP programs and the physical characteristics of most generators makes it relatively straightforward to specify the obligations of a resource that qualifies as ICAP. In general, such resources are required to bid their capability into the ISO's day-ahead energy and ancillary services markets. If the energy is, in real time, being sold to an entity outside the ISO's control area, the ISO has the right to recall that energy and capacity during an emergency. The ICAP programs also have rules that encourage the owners of generation to maintain high availability levels because the capacity credit of each unit is based on its unforced capacity. In other words, frequent forced outages reduce the amount of capacity that qualifies as ICAP. Finally, the ISO has the right to determine when ICAP resources can be taken out of service for planned maintenance.

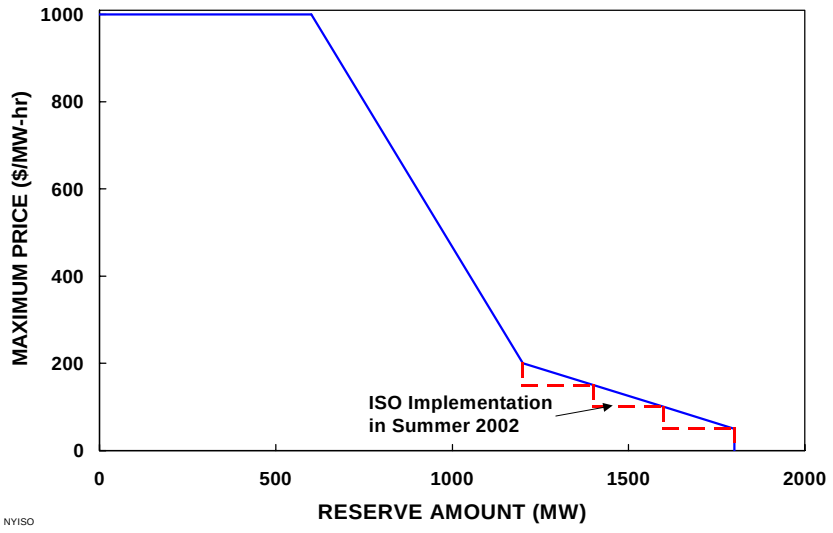
For demand resources, additional complications arise. These complications relate primarily to the energy-limited nature of retail loads. While many customers might be willing to interrupt some of their load for short periods during emergencies, most will not want to endure frequent, long-duration interruptions. The characteristics associated with qualifying

Exhibit 2. Demand Curves for Contingency Reserves

The NERC and regional-reliability-council rules on contingency-reserve requirements are deterministic. That is, they specify minimum amounts of spinning reserve, supplemental reserve (and sometimes 30-minute reserve) as functions of the first and second largest contingencies or expected daily peak demand. The rules imply that these reserve amounts are infinitely valuable, i.e., no price is too high to pay to maintain the required amounts.

These rules are inconsistent with both economic theory and actual operations. Theory suggests that the value of reserves can be no more than the value of lost load. Implicitly, when regulators impose price caps in wholesale markets ($1000/MWh in New York), they are estimating the value of lost load. Thus, in New York, the ISO should pay no more than $1000/MW-hr for any of the reserve services. Control-area operators often run short of reserves when the price of additional resources is extremely high or when the only way to maintain reserves is to involuntarily interrupt some load.

The New York ISO plans to include demand curves for 30-minute reserves, which will let the ISO be deficient in these reserves (but not 10-minute reserves) when the price gets too high. In addition, the ISO will permit gradual restoration of reserves following a reserve pickup (rather than the instantaneous recovery now reflected in the ISO's software). This gradual recovery will prevent post-contingency price spikes from occurring.



Possible demand curve for contingency reserves. The dashed line shows the implicit demand curve used during Summer 2002, based on the New York ISO's treatment of exports.

demand resources are very similar to those for the existing utility and ISO demand-response programs, related to payments, penalties, baseline determination, advance notice, and limits on

the number and duration of interruptions. Perhaps the best way to think of long-term resource adequacy is as an umbrella within which existing and new demand programs can fit.

Key demand characteristics for such programs include:

- Obligation period, the planning horizon (two to four years, annual, seasonal, monthly or daily) over which LSEs must plan for and procure sufficient resources to meet their resource-adequacy obligations. How should demand resources that have strong seasonal characteristics, such as air conditioning, be treated? Such resources are quite valuable during certain seasons (e.g., summer) but provide zero capability during other seasons. How would an LSE certify the demand resources it plans to use to meet its resource-adequacy obligations? Would these certification requirements become increasingly stringent as time passes (from three years out to two years out, and so on until one month ahead of operations)? Requiring commitments a few years ahead of operations provides retail customers with sufficient time to make any capital investments necessary to implement load reductions. On the other hand, many customers will not know their production schedules and other operating limits so far ahead; they can commit to changes in operations much more easily on a week- or day-ahead basis.

- Minimum amount of advance notice from the system operator. For contingency reserves, this time is 10 minutes. The notification time for demand-response programs is often longer, 30 to 120 minutes. These differences complicate payments to the two kinds of resources.

- Minimum amount of time (duration) the resource must remain online delivering energy or maintaining its stated load reduction. For contingency reserves, the reliability rules typically specify a minimum run time of 60 minutes. Demand resources vary enormously in their ability to be interrupted for long times. Residential direct-load-control programs likely offer resources that can be called upon over and over again so long as the interruption period for each event is short (e.g., less than an hour). This is especially true for end uses with storage, such as residential electric water heating. The opposite is true for other end uses. Some industrial processes can be economically shut down only if they remain offline (and are being paid for the interruption) for at least a few hours. This is especially true for situations in which the fixed cost of a shutdown is substantial and can be justified only if the shutdown lasts long enough.[*]

- Frequency with which RTO can call upon the resource. This characteristic applies only to demand resources, which may not be willing to be interrupted too often. Limits on the number of times the demand resource can be called upon each year (or each season)

[*]This situation is analogous to the startup costs associated with generating units.

makes them less valuable to the system operator than supply resources.[*] It is difficult, however, to figure out how much less valuable. If an emergency occurs in June, should the ISO call on the contracted demand resources or should it hold these resources in reserve because more and more-severe emergencies might occur in July and August? These probabilistic questions generally don't arise with supply resources.

- Energy strike price, the price at which the contracted capacity will be converted into energy. For supply resources, this is almost always the real-time spot price of energy. For demand resources, this could be a minimum (floor) price set by the program, the real-time spot price of energy, or zero.[#]

- Provisions to account for the unavoidable less-than-100% availability of the resource. Should resource owners be obligated to inform the system operator whenever the resource is unavailable and for how long the outage is expected to last? How might such a reporting requirement apply to participating demand resources? How should the RTO account for the fact that the aggregate load reduction from many retail customers (each of which provides only a small load reduction) is more reliable than the reserves obtained from a few large generators?

What types of demand resources should qualify for long-term resource adequacy? FERC is clear that biddable demand (which is not defined) and interruptible demand qualify. Should the LSE be required to specify, three years ahead of operations, the processes and equipment that will provide the necessary load reductions. Should all resources be subject to certification tests, in which the amount of capacity provided or shed is demonstrated?

Should energy-efficiency programs and measures qualify as resources, or should they be treated as reductions in the load forecast for the relevant LSEs? Because energy-efficiency measures (e.g., efficient ballasts for fluorescent lamps) are passive rather than dynamic, system operators cannot call on them as resources in real time and they don't meet the security requirements of NERC's reliability definition. That is, they operate much as large nuclear, baseload power units. On the other hand, by lowering demand, they reduce the need for new generating units and transmission facilities. My preliminary recommendation is that these efficiency improvements should be incorporated in the long-term load forecasts that determine how much resources are required. In this scheme, energy efficiency would not be considered a qualifying resource and would get no explicit credit or payment, but its existence will reduce the overall resource requirement and therefore save money for the LSE (Exhibit 3).

---

[*]Pacific Gas & Electric exhausted its rights to the 500 MW of interruptible load it had early in 2001, during the height of the California electricity crisis (California Public Utilities Commission 2001). As a consequence, these resources were not available to the utility or the ISO for much of that year.

[#]Ruff (2002a) suggests that retail loads should be paid nothing for load reductions because they already benefit from not having to pay the high price for the power they are not consuming. To pay them for load reductions would be to pay them twice, an inappropriate subsidy.

Exhibit 3. Treating Demand Response as Load Reductions or Resources

Treating demand resources as reductions in forecast load vs treating them explicitly as qualifying resources makes a big difference in their value. Demand resources count for more as load reductions.

Consider an LSE with a 100 MW peak load and a requirement to acquire 115 MW of capacity for its resource-adequacy obligation, based on a 15% planning-reserve margin. Assume this LSE has 15 MW of demand response (e.g., industrial interruptible load and direct control of residential water heaters) that it can call on. If this response is treated as a resource, the LSE then has another 100 MW of resource requirement to fill, presumably with generating units. However, if the 15 MW of demand resources is considered a load reduction, it now needs only 98 MW of additional resources [(100 – 15)*1.15]. Thus, demand response is more valuable as a reduction in load than directly as a resource.

How about price-responsive demand? Would it be enough for an LSE to say to the RTO "I have x MW of demand that will bid into the day-ahead market* such that at the $1000/MWh price cap zero MW will be purchased"? My preliminary recommendation is the same as the one for energy efficiency. The price-responsive demand should show up in the forecast of peak demand that determines the Resource requirements for each LSE. Thus, price-responsive demand would, once again, receive no explicit credit or payment, but its existence would reduce the overall requirement and therefore save money for the LSE.

How might the locational benefits of demand resources be captured? These benefits include those associated with lower (or zero) transmission losses and congestion. In New York, because of its severe transmission constraints, the UCAP requirements are locational. In general, a resource seeking qualification might be required to demonstrate that the output from that resource can be delivered to the control area (or perhaps to a particular zone within the control area).

What is the tradeoff, for various kinds of participating customers, between the upfront reservation (capacity) payment and the performance (real-time energy) payment? Customer surveys show that retail customers prefer reservation payments that are independent of the frequency and severity of emergencies to highly uncertain real-time payments (i.e., they prefer capacity payments to energy payments). On the other hand, these same customers strongly dislike penalties, which are the quid pro quo for capacity payments (Rosenstock 2001). That is, if the LSE or the ISO pays for capacity upfront, it is purchasing the right to convert that capacity into energy reductions under certain conditions. Failure to perform on the part of the

*Must it also bid these price-sensitive demands into the real-time market?

retail customer should result in a penalty, not just forfeiture of the capacity payment. Further, these capacity payments and nonperformance penalties should be consistent with those that apply to supply resources.

Given the existence of several kinds of demand-response programs, what is the sequence in which the ISO should call on these resources and programs? This sequence could be based on the existence and amount of an upfront payment, with resources that are paid a reservation fee called before those that do not receive such payments. Also, those resources that require more upfront notification could be called sooner than those with a shorter notification period. This is an important issue for demand resources because of the proliferation, during the past few years, of ISO economic and emergency demand-response programs. A related question is whether the same demand resource can participate in more than one program. Is this a form of double dipping, and is it inconsistent with how generators are treated in wholesale electricity markets?

Does the baseline need to be specified at the time the LSE declares the resource (several years ahead)? Similarly, what metering, communications, computing, and controls infrastructure is required to meet the resource-adequacy requirements, and must these infrastructure details be specified several years in advance?

Perhaps the solution is to require qualifying demand resources to bid daily into reserve markets. They can bid a very high energy price but they are then available to the RTO. Ruff (2002b) argues that paying a reservation charge to retail loads that participate in ancillary-service markets is appropriate because ancillary services are a different set of products from energy:

> A generator providing reserve service should be paid for standing by to be ready to provide energy when told to do so by the ISO, and then should be paid the market price for the energy it provides. Similarly, a load providing reserve should be paid for being ready to reduce load when told to do so by the ISO, and then should pay the market price for the energy it actually takes, i.e., it should save the market price for the energy it does not take. The ancillary services provided and paid for in this case is the service of being READY to change market sales/purchases and doing so when told to do so by the ISO, not actually providing/not-taking the energy. The market energy price provides all the compensation that is needed or appropriate for the energy actually delivered/not taken.

Also, the baseline for load reductions is much easier to determine with contingency reserves because the notification period is only 10 or 30 minutes.

5. ISO ICAP PROGRAMS

The three northeastern ISOs (PJM, New York, and New England) have had in place long-term resource-adequacy requirements for many years. PJM and New York also operate markets for the purchase and sale of installed capability. PJM-West has a shorter-term available capability (ACAP) requirement, and various groups in California are developing ACAP proposals that the California ISO might implement as part of its Market Design 2002.

PJM

As part of its ICAP program, PJM (2002a) has operated an Active Load Management (ALM) program since 1991. This program, operated primarily by the distribution utilities, includes direct control of residential equipment, customer load reduction to a firm level (interruptible contracts), and guaranteed load drops implemented through the use of onsite generation. In this program, PJM provides no monetary payment. Instead, participating utilities receive installed-capability credits for the load reductions, which reduce the utilities' costs of installed generating capacity. Participating loads must be available for up to 10 PJM-initiated interruptions during the planning period (October through May and June through September), for interruptions lasting up to six hours between noon and 8 pm on weekdays, and within one or two hours of notification to the LSE by PJM.[*] PJM has, in this program, no direct contact with the retail customers; PJM deals with the LSEs, and the LSEs deal, in turn, with the customers. In particular, any payments to loads for their agreement to curtail during system emergencies is arranged between the LSE and its customers. Such arrangements do not involve PJM, but are under state regulatory authority.[#]

The baseline against which load reductions are measured is either the customer's load one hour before the event or the customer's hourly loads on a comparable day, as determined by the LSE. Failure to perform can lead to penalty charges related to PJM's capacity deficiency charge; that is, the penalty is comparable to that which would apply for providing insufficient generating capacity to meet the required installed-capability requirement. The penalty is equal to the Daily Capacity Deficiency Rate multiplied by 365/10. The deficiency rate is based on the annualized cost of a new CT; the factor of 10 is based on PJM's right to call on these resources up to 10 times a summer. The penalty is set very high (10% of the annual fixed costs of a CT) to encourage compliance with PJM interruption requests.

PJM does not treat ALM as an ICAP resource; rather the ALM lowers the peak load the LSE is responsible for. The load reduction is not a one-for-one reduction in ICAP obligation:

---

[*] During an emergency, PJM calls on the slower-responding (2-hour advance notice) resources before it calls on the 1-hour resources.

[#] For example, Baltimore Gas & Electric has 283 MW of load-response resources, most of it in residential air conditioning. When emergencies occur, PJM contacts BG&E, which then dispatches these demand resources.

Peak load - (ALM Credit × ALM factor)×Forecast Pool Multiplier

where the ALM factor is 0.966, adjusted down from 1.0 because it applies to only the four summer months; the Forecast Pool Multiplier is based on the installed reserve margin and poolwide equivalent forced outage rate and is equal to 1.0897. The ALM amount accounts for transmission and distribution losses, by building back up to the generator level from the customer level.

Until recently, about 1,700 MW of load (roughly half of which is residential and small-commercial direct-load control and half of which is industrial loads and onsite generation) qualify for installed capability in PJM. The program was called upon six times during the summer of 1999 but not at all during the summer of 2000. About 600 MW of the ALM load migrated to other PJM programs in 2002, leaving about 1275 MW that year.

The sequence PJM uses to call upon resources is maximum emergency generation and then the demand programs, first the demand-response programs, then ALM, then voltage reduction. In practice, programs are called at the same time.

Although PJM operates markets for ICAP, ALM is not part of these markets. That is, PJM does not permit trading of ALM credits.

PJM-WEST

PJM West was created when Allegheny Energy joined PJM in early 2002. Instead of ICAP, PJM-West uses a shorter-term reserve requirement called ACAP. ACAP applies day-ahead (rather than for multimonth periods as in PJM) and is equal to peak load plus 6% (4% for the ECAR operating reserve requirement plus 2%). This “extra” capacity must be available all 24 hours. PJM runs daily and longer-term markets for ACAP.

Although ACAP allows for demand resources, none are now participating. Qualified Interruptible Load (QIL) includes two classes of load reduction, cycling and load drop (PJM 2002b). QIL must be available for at least five PJM-initiated interruptions during each planning period, must be able to be implemented within two hours of notification, and must be available for interruptions for up to six hours between 7 am and 10 pm.

NEW YORK ISO

The New York ISO imposes installed capability requirements on LSEs and operates capability markets, much as PJM does. Interruptible load can qualify as unforced capacity in New York under what the ISO calls Special Case Resources (New York ISO 2002). The New York markets differ from those in PJM in that SCRs can participate in the New York capability markets.

SCRs must provide a minimum of 0.1 MW of load interruption. About 300 MW of load reduction and onsite generation qualified as SCR in 2001, and about 600 MW qualified in 2002. The sponsoring LSEs are given a day-ahead notice that an emergency might occur (e.g., the ISO anticipates a shortage of contingency reserves) and then an in-day 2-hour notice for deployment (i.e., load interruption). The actual load reduction is increased by the Transmission District loss factor to make the UCAP amount comparable to that provided by generating units.

SCRs are exempt from some of the ISO's reporting and certification requirements that apply to other UCAP resources: "Special case resources are not subject to bidding, scheduling, and notification requirements." An SCR "may be required by the ISO to demonstrate its pledged load reduction capability once in every capability period if it has not otherwise already been called upon by the ISO to reduce load in such period." SCRs receive the same upfront reservation payments as do other resources and receive no performance payments. As with other UCAP resources, failure to comply with an ISO request can lead to penalties. Retail loads qualifying as SCR may also participate in the ISO's other demand-response programs, the Emergency Demand Response Program or the Day-Ahead Demand Response Program.

The New York ISO is contemplating changes in its demand-response programs. In particular, it is thinking of calling the SCRs first when emergencies occur, primarily because these resources, unlike those participating in the other demand-response programs, receive upfront payments. The ISO is also thinking of permitting SCRs to set an energy strike price so that the ISO can better integrate the supply and demand sides of its real-time energy market in establishing a real-time energy price that accurately reflects the current physical conditions in the grid.

ISO NEW ENGLAND

ISO New England, like the two other northeastern ISOs, requires LSEs to acquire or own sufficient resources to meet planning-reserve goals. Unlike the two other ISOs, New England does not operate markets for installed capability.

New England, however, plans to adopt the New York approach to capability requirements. ISO New England's (2002b) proposed *Market Rule 1* and FERC's (2002) approval of New England's Standard Market design discuss ICAP (i.e., Section 8 of Market Rule 1). The proposal includes both summer and winter seasonal ICAP requirements based on unforced capacity. The ISO will conduct monthly UCAP auctions and will impose a deficiency charge of $6,660/MW-month on LSEs that are deficient. Unlike New York, New England will not develop locational ICAP requirements. However, both ISO New England and FERC are silent on whether and how demand resources can qualify for unforced capability.

CALIFORNIA ISO

The California ISO (2002), as part of its Market Design 2002 proposal to FERC, proposes to institute an ACAP requirement, which would operate on a monthly and daily basis. The focus of the California proposal is on *available* rather than *installed* capacity to adjust for forced and maintenance outages. LSEs will be required to identify a month ahead what resources they will draw on to meet their loads and contingency-reserve obligations. At some point, the responsibility for available capacity shifts from the LSE to its suppliers, which then are responsible for any deficiencies associated with a forced outage. Thus, ACAP requires 100% availability day ahead and in real time. The ISO does not plan to run ACAP market, but ACAP resources must bid into day-ahead and real-time markets. The ACAP requirement will likely equal 112% of estimated monthly peak demand.

Although the ISO offers many details concerning the objectives of this requirement, the obligations of the buyers and sellers of ACAP, the definition of reserve margin, the load-forecasting responsibilities, allowable resources, location-specific requirements, planning horizon, and penalties, the proposal is silent on the role of demand resources. Conversations with ISO staff suggest that demand resources will be encouraged to participate fully in ACAP, but the details are not yet worked out.

6. CONCLUSIONS

Existing wholesale spot markets are far from perfect. These imperfections—including inefficient market designs and rules, potential abuse of market power by generators, and bid and price caps—prevent day-ahead and real-time prices from accurately reflecting the costs and value of electricity. As a consequence, potential investors in new generation and demand-management resources may have insufficient incentives to make such investments.

Because of these problems, FERC and others suggest that all load-serving entities be obligated to demonstrate the rights to sufficient resources to meet the peak demands of their customers plus a margin to provide for contingency reserves, planned outages, forced outages, and other unexpected events. The three northeastern ISOs have all run such programs during the past few years focusing on unforced capability (installed capability that is derated to account for forced outages). FERC proposed a comparable long-term resource-adequacy requirement in its recent notice on standard market design.

These programs and proposals help to ensure that enough generation, transmission, and demand-management capacity will exist to reliably meet customer electricity needs. On the other hand, these programs and proposals raise many questions about resource definition and program implementation. To a large extent, these issues have not yet been adequately resolved.

Designing a suitable role for demand resources within resource-adequacy programs and requirements is difficult until the underlying issues are resolved. The key principle is simple to state: demand resources should participate on an equal basis with supply resources. That is, it should be subject to comparable, but not necessarily identical, requirements for response, metering, payments, and penalties. Meeting these requirements will be challenging because demand resources are not identical to supply resources in many characteristics. One useful way to view the role of demand resources in such long-term efforts is as a box in which existing ISO and utility demand-response programs and pricing approaches can qualify.

REFERENCES

J. G. Besser, J. G. Farr, and S. F. Tierney 2002, "The Political Economy of Long-Term Generation Adequacy: Why an ICAP Mechanism is Needed as Part of Standard Market Design," *The Electricity Journal* **15**(7), 53–62, August/September.

California Energy Commission 2002, *2002–2012 Electricity Outlook Report*, Sacramento, CA, February.

California Independent System Operator Corp. 2002, Letter Transmitting Proposed Tariff Revisions, from C.F. Robinson et al., Docket Nos. ER02-1656-000 et al., Folsom, CA, June 17.

California Public Utilities Commission 2001, *Energy Division's Report on Interruptible Programs and Rotating Outages*, San Francisco, CA, February 8.

E. Hirst and S. Hadley 1999, *Maintaining Generation Adequacy in a Restructuring U.S. Electricity Industry*, ORNL/CON-472, Oak Ridge National Laboratory, Oak Ridge, TN, October.

ISO New England 2002a, *Request for Approval of Revisions to NEPOOL Market Rules 1, 2, 3, 4, 5, 6, 8, 9, 15, and 17*, Docket Nos. ER02-1149-000 and -001, submitted to the Federal Energy Regulatory Commission, Holyoke, MA, February 27.

ISO New England 2002b, *Market Rule 1*, Holyoke, MA, July 15.

New York Independent System Operator 2002, *NYISO Installed Capacity Manual*, Schenectady, NY, August 30.

North American Electric Reliability Council 2002, *Reliability Assessment 2002–2011*, Princeton, NJ, October.

D. B. Patton 2001, *An Assessment of Peak Energy Pricing in New England During Summer 2001*, Independent Market Advisor to ISO New England, Holyoke, MA, November.

PJM Interconnection, LLC 2002a, *PJM Manual for Load Data System, Manual -19, Revision: 04*, Norristown, PA, June 1.

PJM Interconnection, LLC 2002b, *PJM West Capacity Market Business Rules*, Norristown, PA, June 25.

S. Rosenstock 2001, "Views on Demand Response Programs from a National Accounts Customer," Peak Load Management Association Conference, November 8.

L. E. Ruff 2002a, *Economic Principles of Demand Response in Electricity*, Edison Electric Institute, Washington, DC, September 3.

L. E. Ruff 2002b, Personal Communication, October.

R. J. Shanker 2002, "Is an Adequacy Market Necessary?" Presentation to Electric Power Generation Association, Washington, DC, October 16.

S. Stoft 2002, *Power System Economics: Designing Markets for Electricity*, IEEE Press, Wiley-Interscience, New York, NY.

U.S. Federal Energy Regulatory Commission 2002a, *ISO New England, Inc., Order Accepting Amendments*, Docket Nos. ER02-1149-000 and -001, Washington, DC, April 26.

U.S. Federal Energy Regulatory Commission 2002b, *Notice of Proposed Rulemaking: Remedying Undue Discrimination through Open Access Transmission Service and Standard Electricity Market Design*, Docket No. RM01-12-000, Washington, DC, July 31.

U.S. Federal Energy Regulatory Commission 2002c, *Order Accepting in Part and Modifying In Part Standard Market Design Filing and Dismissing Compliance Filing*, Docket Nos. ER02-2330-000, Washington, DC, September 20.

C:\Data\Docs\NEDRI\LTResourceAdequacyDraft.wpd