PRICE-RESPONSIVE DEMAND AS RELIABILITY RESOURCES*

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

April 2002

1. INTRODUCTION

Permitting and encouraging at least some retail consumers to face time-varying electricity prices offers three sets of benefits: economic, environmental, and reliability (Hirst and Kirby 2001). This paper focuses on the reliability benefits of price-responsive demand.#

Bulk-power systems are fundamentally different from other large infrastructure systems, such as air-traffic control centers, natural-gas pipelines, and long-distance telephone networks. Electric systems have two unique characteristics:

- The need for continuous and near instantaneous balancing of generation and load, consistent with transmission-network constraints: this requires metering, computing, telecommunications, and control equipment to (1) monitor loads, generation, and the transmission system, and (2) adjust generation output to match load.

- The transmission network is primarily passive, with few "control valves" or "booster pumps" to regulate electrical flows on individual lines; control actions are limited primarily to adjusting generation output and to opening and closing switches to add or remove transmission lines from service.

Because of these two characteristics, bulk-power system operators rely primarily on changes in generation output (MW movements up or down) to keep the system in balance and to comply with transmission limits. In principle, changes in electricity consumption could serve as well as generator movements in meeting these reliability requirements, but the use of customer loads for reliability purposes is the exception rather than the rule.

The traditional, vertically integrated utility managed short-term reliability by dispatching its own generating units as well as adjusting transformer settings and turning breakers on and off at its own transmission facilities. In competitive wholesale markets, system operators

*Preparation of this paper was supported, in part, by the Regulatory Assistance Project, Montpelier, VT.

#See publications from the U.S. Federal Energy Regulatory Commission [FERC; Lafferty et al. (2002)] and Braithwait and Faruqui (2001) for discussions of the economic benefits of price-responsive demand programs; Cowart (2001) addresses the reliability and environmental benefits of such programs.

increasingly work for entities that own no generation and, in many cases, own no transmission. In such cases, the system operator must establish markets for the generation reliability services and negotiate contracts with transmission owners for other reliability services.

This change in industry structure and the associated emergence of wholesale energy and reliability markets create new opportunities for demand-side resources. If the market rules are technology neutral (i.e., they focus on the function to be performed and not on how that function is done), customer loads will be able to participate in these markets. Such participation will either enhance reliability or lower the costs of maintaining reliability for all customers (by deepening reliability markets) and will save money for participating customers.

The discussion so far has focused on short-term operations, what system operators must do in real-time and near-real time (e.g., up to day ahead) to maintain reliability. The North American Electric Reliability Council (NERC) calls these functions *security*. Long-term planning, design, and investments involve what NERC calls *adequacy*.[*] Adequacy requires the installation of sufficient generation and transmission resources to meet reasonably foreseeable consumer electricity demands. Here, too, customer loads should be able to contribute to reliability by making long-term commitments to reduce loads during emergency conditions.

The next section of this paper discusses current North American reliability rules and shows how these requirements discriminate against loads in the provision of reliability services. Section 3 explains the wholesale markets that today's independent system operators (ISOs) run and that tomorrow's regional transmission organizations [RTOs; see FERC (1999)] will likely manage. These markets include day-ahead and real-time energy and congestion management, ancillary services, and long-term markets for installed capability and transmission congestion. Section 4 proposes a new way to treat involuntary load reductions, for both equity and efficiency reasons. The final section summarizes the findings from this paper and offers suggestions on the use of retail loads to provide bulk-power reliability services.

2. RELIABILITY RULES AND PRACTICES

Responding to a major generation outage provides an example of how the electricity industry responds to its unique features. Figure 1 illustrates how the electric system operates when a major generating unit suddenly fails. Prior to the outage, system frequency is very close to its 60-Hz (cycles per second) reference value. Generally, within a second after the outage

[*]NERC defines reliability as "the degree to which the performance of the elements of [the electrical] system results in power being delivered to consumers within accepted standards and in the amount desired." NERC's definition of reliability 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." It 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.

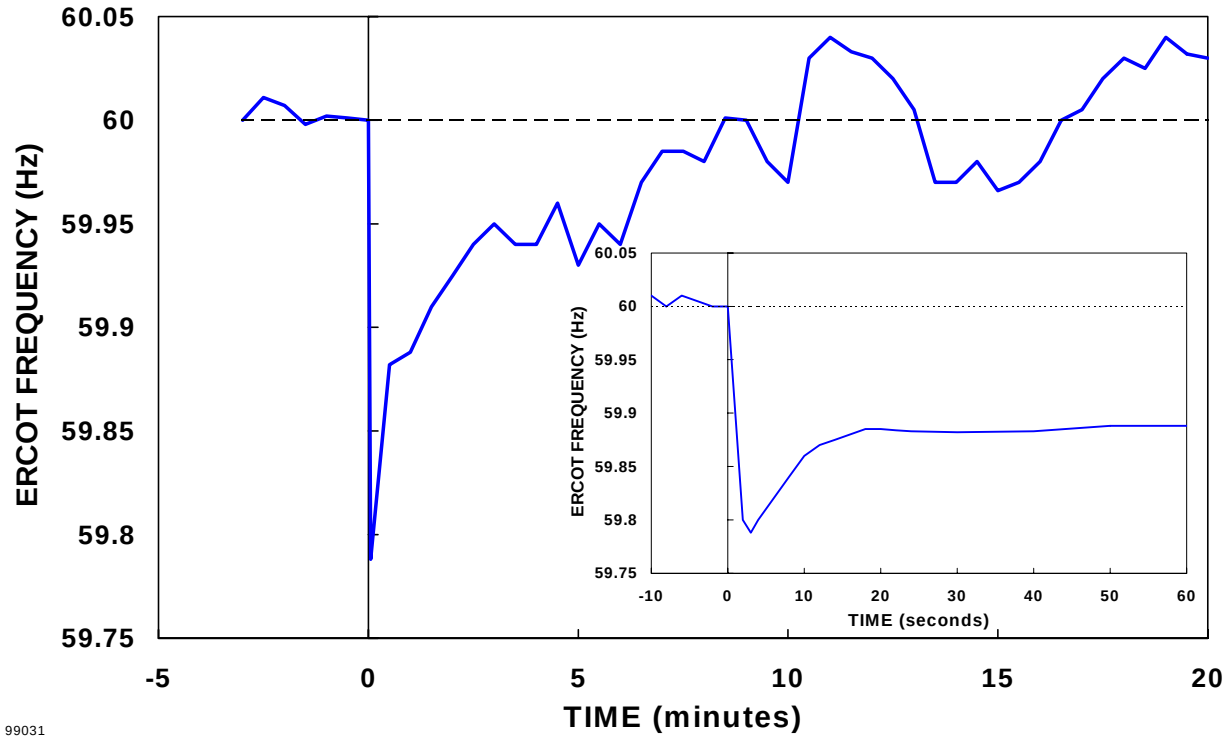

Fig. 1. **Interconnection frequency before and after the loss of a 653-MW generator. The inset shows frequency for the first minute after the outage, and the larger figure shows frequency for the first 20 minutes after the outage.**

occurs, frequency drops, in this case to 59.79 Hz. The frequency decline is arrested primarily because many electrical loads (such as motors) vary with system frequency. If the frequency decline is large enough, the governors at those generators so equipped sense the frequency decline and open valves on the turbines, which rapidly increases generator output. This governor response accounts for the initial increase in frequency during the first several seconds after the outage occurs, as shown in the Fig. 1 inset. At this point, the generating units that provide contingency reserves, in response to signals from the control center, increase output. In this example, the system worked as it was intended to, and frequency was restored to its pre-contingency 60-Hz reference value within the required 10 minutes (at 8.5 minutes).*

A reasonable question to ask, in reviewing this example and Fig. 1, is whether this reliability problem could have been solved, at least in part, by a reduction in load. The answer, of course, is yes—in principle. I write "in principle" because in practice customer loads are often not allowed to participate with generation on an equal footing in provision of reliability services.

NERC, established by the electric-utility industry in 1968, develops standards, guidelines, and criteria for assuring system security and evaluating system adequacy. Existing

*In early 2000, NERC extended the allowable disturbance-recovery period from 10 to 15 minutes.

NERC Policies inappropriately favor generation resources over customer loads in the provision of reliability (ancillary) services. Consider, as an example, NERC's (2001c) Policy 1 — Generation Control and Performance.* This policy deals with the generation:load balance required under normal operating conditions and under emergency (contingency) conditions. The Policy refers to three kinds of reserves used to respond to a major contingency (e.g., loss of a large generator or major transmission line): frequency response, spinning reserve, and supplemental reserve. The discussion of frequency response deals only with generating units with governors (e.g., governor droop, deadband, and limits). No mention is made of loads providing frequency response.

Policy 1 defines spinning reserve as "unloaded generation that is synchronized and ready to serve additional demand." Clearly, this statement excludes customer loads from providing this valuable and expensive ancillary service. NERC's definition of nonspinning reserve, on the other hand, does allow for the use of loads to provide this service: "that operating reserve not connected to the system but capable of serving demand within a specified time, or interruptible load that can be removed from the system in a specified time." Unfortunately, NERC's definition of interruptible load is rather narrow: "demand that can be interrupted by direct action of the supplying system's system operator in accordance with contractual provisions."

These distinctions affect economic efficiency in two ways. First, NERC's Policy 1 currently requires that "at least 50% of operating reserve shall be Spinning Reserve." Second, spinning reserve is much more expensive than nonspinning reserve. If loads were permitted to supply spinning reserve, they would have more incentive to participate in markets for ancillary services and the prices for spinning reserves would decline. This change would enhance reliability by providing more resources for contingency reserves and would save money for electricity consumers by lowering the costs for these reserves.

In New York, the price of spinning reserve was, on average, 50% higher than the price of nonspinning reserve ($3.0 vs $2.0/MW-hr) over the 20-month period from April 2000 through November 2001 (Fig. 2). In the Electric Reliability Council of Texas, the price of responsive (spinning) reserve averaged $6.4/MW-hr from September through November 2001, compared with only $1.4/MW-hr for nonspinning reserve. Finally, the prices for spinning and nonspinning reserves in California for 1999 averaged $6.5 and $3.6/MW-hr, respectively. Clearly, spinning reserve is a more valuable and, therefore, more expensive service. Why should the demand side be precluded from participating in these lucrative markets? More important, why should certain resources be prohibited from performing these valuable reliability functions?

*Note that this policy deals with *generation* and not *demand*.

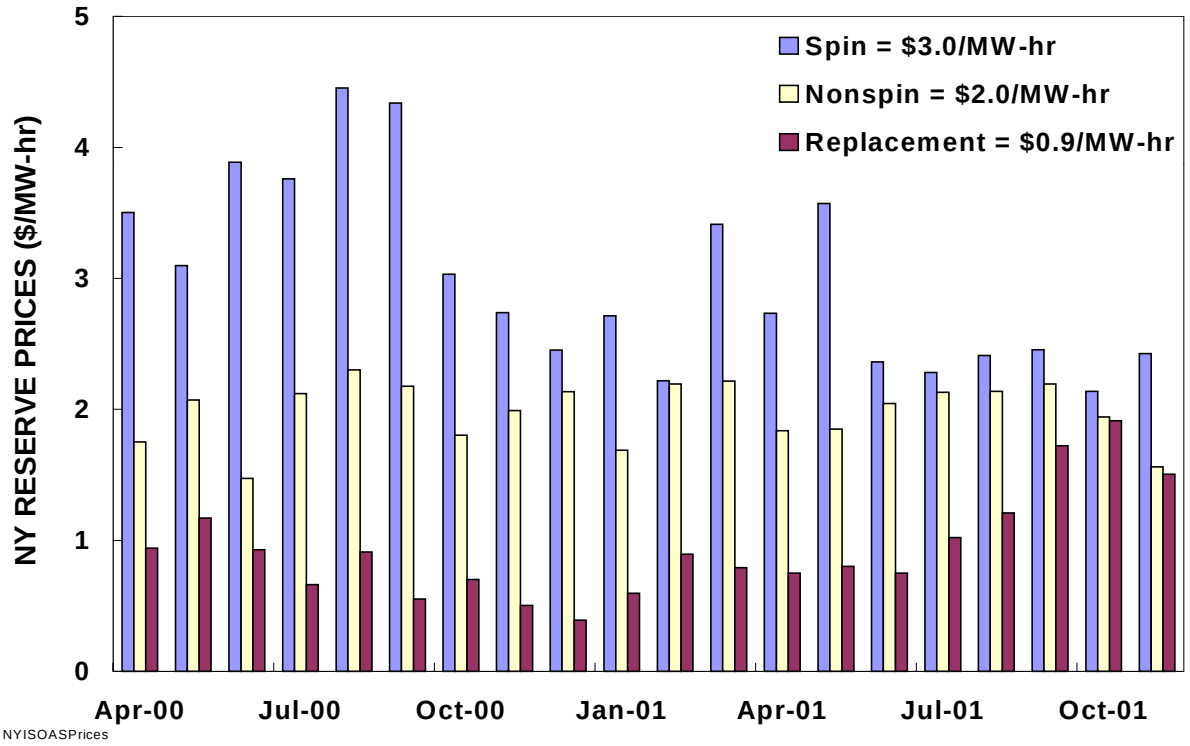

Fig. 2. Prices for operating reserves in New York from April 2000 through November 2001.

Fortunately, NERC (2001b) recognizes these limitations in its current operating policies. Its draft Policy 1 defines spinning reserve as "the Resource Capacity in excess of current and anticipated demand that is synchronized to the grid and deployable" including "controllable load resources." Note that this definition is technology neutral; it does not specify whether the required capacity comes from generation or customer loads, a welcome modification.

In a similar fashion, the proposed Policy 1 defines a new Frequency Response Standard with Frequency Responsive Resources that "may be any combination of natural load response, generator governor action, under-frequency load shedding of contractual interruptible load …, or any other equipment that meets the desired characteristics." Once again, the proposal—but not the current policy—is technology neutral, permitting customer loads to provide reliability services on an equal footing with generating resources. More generally, NERC's (2002) new reliability model calls for standards that neither favor nor preclude any market mechanism or technology.

In 1993, NERC (1993) issued a reference document that explains how customers loads can substitute for some reliability services traditionally provided by generators, especially operating reserves. This paper distinguishes between indirect (passive) and direct (active) load management, with the latter involving the ability of the system operator to take action to disconnect the load. The extent to which NERC considers changes in customer loads a

reliability resource depends on the ability of the system operator to control that load change, either directly through automatic controls (e.g., to turn off electric water heaters) or indirectly (e.g., through a phone call to a facilities manager). In addition, this document suggests that the system operator must know what the load is both before and after the exercise of control actions.

This statement on the information that must be made available to the system operator appears to require that (1) each controlled load be individually metered at, say, the 1-minute level and (2) the load's electricity consumption be telemetered to the control center in real time. An alternative would permit aggregation of enough loads so that the system operator need see only the combined response of all the loads. Such aggregation is especially important for residential loads that might provide contingency reserves because it might be too expensive to install interval meters and the associated communications systems on individual appliances. A single residential water heater might provide only about 0.0005 MW of load reduction. However, an aggregation of 100,000 water heaters could provide 50 MW of load relief, enough to be visible to the system operator without any special metering or communications equipment.

This NERC document suggests that loads must be interruptible within 20 cycles (⅓ second) to qualify for spinning reserve. This is a surprising requirement for two reasons. First, NERC's Operating Policy 1 imposes no such requirement on generators that provide spinning reserve. Second, this requirement is not needed to comply with NERC's Disturbance Control Standard (DCS), which requires only that recovery be completed within 15 minutes. Rather, this 20-cycle requirement seems to be related to a separate service, frequency response, that is not yet part of the NERC reliability system.

3. WHOLESALE ELECTRICITY MARKETS

Wholesale electricity markets typically include long-term markets for transmission rights (either financial or physical) and installed generation capability as well as short-term markets for energy, ancillary services, and congestion (Chandley 2001). This section focuses on the short-term markets, which operate between day ahead and real time (Table 1).

Although the explanations in Table 1 emphasize the role of generators, rather than loads, there is no theoretical reasons why loads cannot provide any of the services. This section discusses these services and explains how loads could participate in the relevant markets.

Table 1. Wholesale markets for energy, ancillary services, and transmission congestion

Market	Description
Day-ahead energy and congestion management	Potential suppliers submit bids for hourly energy ($/MWh), and perhaps also startup ($) and no-load costs ($/hr). These resources are used to balance generation and load on an hourly basis, respecting all transmission constraints as scheduled day ahead
Day-ahead ancillary services	Potential suppliers submit capacity ($/MW-hr) and energy ($/MWh) bids to supply service on an hourly basis. These resources are required to meet NERC security requirements in real time
Regulation[a]	Generators online, on automatic generation control, that can respond rapidly to system-operator requests for up and down movements; used to track the minute-to-minute fluctuations in system load and to correct for unintended fluctuations in generator output to comply with NERC's Control Performance Standard
Spinning reserve[a]	Generators online, synchronized to the grid, that can increase output immediately in response to a major generator or transmission outage and can reach full output within 15 minutes to comply with NERC's DCS
Supplemental reserve[a]	Same as spinning reserve, but need not respond *immediately*, therefore units can be offline but still must be capable of reaching full output within the required time
Replacement reserve	Same as supplemental reserve, but with a 30- or 60-minute response time, used to restore spinning and supplemental reserves to their precontingency status
Frequency response	Not "officially" recognized as a service by either NERC or FERC, this service includes generators equipped with governors (and loads) that change output (consumption) in response to changes in Interconnection frequency
Real-time energy and congestion management[b]	Potential suppliers submit bids ($/MWh) for the amount of energy that can be provided, either up or down, in each interval (5, 10, or 15 minutes). These resources are used in real-time to maintain the necessary generation:load balance and to relieve transmission congestion

[a]FERC (1996) included these three ancillary services in its Order-888 open-access transmission requirements.

[b]The real-time (intrahour) energy market provides the service that FERC called energy imbalance in Order 888.

DAY-AHEAD ENERGY MARKET

Although most electricity is bought and sold under long-term bilateral contracts, perhaps 5 to 10% will be traded either day-ahead or in real-time. These short-term trades could be a consequence of changed circumstances (e.g., a competitive retail provider signed up more customers than it anticipated, a generator completed its planned maintenance outage faster than it expected to, a generator suddenly trips offline, or the weather is different from what was expected) or they could be part of a company's risk-management strategy (maintaining a portfolio of long- and short-term options). In any case, the typical RTO design, as exemplified by those in PJM and New York, calls for suppliers and buyers to submit hourly bids by, say 10 am, on the day before the operating day.[*] The RTO then evaluates these bids using its security-constrained unit-commitment optimization computer model (Hirst 2001). This model schedules generation and price-responsive demand hour-by-hour for the operating day so as to respect all generator and security (reliability) constraints and to minimize operating costs. The output from this model is a set of prices that vary from hour to hour and location to location (either by node or zone) and the schedules of output and consumption for each successful bidder in this day-ahead market. In such a system, demand and supply compete on an equal footing and are fully integrated in the unit-commitment process.[#]

Most such markets today operate with virtually no price-responsive demand. That is, the model treats demand as a fixed quantity each hour and optimizes across the generator bids only. Figure 3 illustrates the two situations. The solid curve that slopes up to the right represents the combined bids from all the generators, showing the typical increase in bid prices associated with increased output. The vertical dashed line reflects the situation that typically occurs today, in which customer demand is specified each hour independent of the price of electricity. The dashed curve that slopes up to the left represents the combined bids from all retail loads that would occur in a fully functioning wholesale day-ahead electricity market. In this case, retail customers or their retail-service providers would specify how much electricity they would use at various prices. In both demand cases, the intersection of the supply and demand curves determines the amount of electricity to be produced and consumed each hour (ignoring losses) as well as the market-clearing price for that hour. These prices and quantities represent binding financial contracts. Any production or consumption of electricity in real time that differs from these final day-ahead schedules is settled at the real-time price.

---

[*]These markets typically permit suppliers to offer three-part bids: (1) the cost to start the unit, (2) the cost to operate the unit each hour absent any electricity production, and (3) a series of bids showing the energy price as a function of unit output, with these bid prices increasing with output level.

[#]Some ISO demand programs permit retail loads to respond to day-ahead or real-time prices, but loads cannot *set* these prices because their price responsiveness is evaluated after the unit-commitment process is completed (rather than integrated into the process).

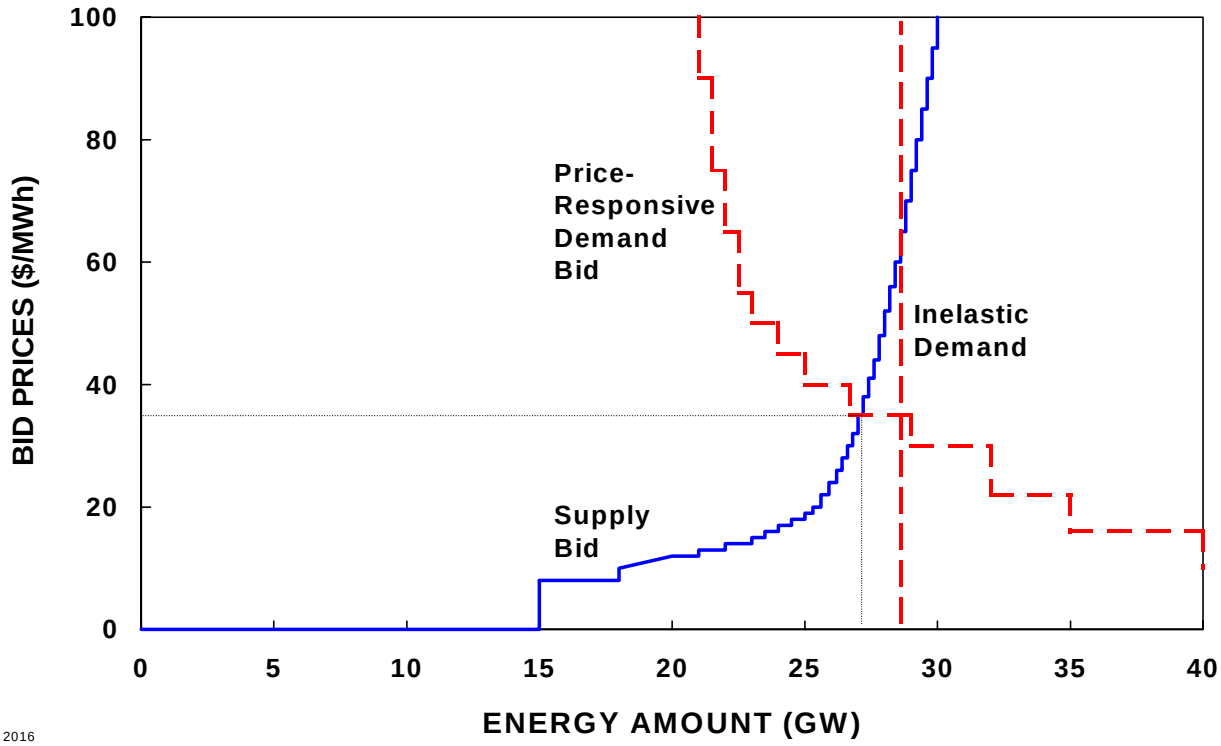

Fig. 3. **Hypothetical generation-supply curve and two demand curves. The vertical demand curve reflects the typical situation in today's wholesale markets, in which retail demand is fixed (independent of electricity price). The dashed line represents the situation in which retail customers bid in different amounts of load as a function of the price.**

Why, one might wonder, are we discussing the day-ahead energy market in a paper on reliability? Although the addition of price-responsive demand to the day-ahead market has strong positive implications for economic efficiency (including a reduction in the ability of generators to exercise market power), it also has reliability benefits. To the extent demand is reduced during high-priced periods, reliability is improved because additional supplies are available to meet any contingencies that might occur. And in efficient markets, prices are high when reliability is threatened. The hypothetical example in Fig. 3 shows a reduction in demand from 29 GW with inelastic demand to 27 GW with price-responsive demand, a release of 2 GW of generating capacity that can be called upon in an emergency.

In addition to freeing up capacity, suitably located demand reductions bid into the day-ahead market can relieve potential congestion problems. The example of Fig. 3 could be repeated for the various zones within an RTO's control area. Depending on the bids and the status of the transmission network, the prices could be quite different across zones reflecting the amount of congestion that would otherwise occur.

ANCILLARY SERVICES

In addition to operating a day-ahead energy market, ISOs also run markets for certain real-power ancillary services (Table 1).* These services are required to respond to the two unique characteristics of bulk-power electric systems noted above, the need to maintain generation:load balance in near-real-time and the need to redispatch generation (or load) to manage power flows through individual transmission facilities.

NERC's Policy 1 on "Generation Control and Performance"specifies two standards that control areas must meet to maintain reliability in real time. The Control Performance Standard covers normal operations and the DCS covers recovery from major generator or transmission outages. The regulation ancillary service is the primary resource system operators use to meet the Control Performance Standards. In principle, customer loads could provide the service as well as generators. Because provision of this service requires a change in output (or consumption) on a minute-to-minute basis and, therefore, requires special automatic generation control equipment at the generator (or customer facility), it seems unlikely that many retail loads will be able to or want to provide this service. Therefore, we do not discuss regulation further.

The three reserve services listed in Table 1 are all intended to help control-area operators meet the DCS. Briefly, DCS requires that the system recovers from a major outage within 15 minutes. The three reserve services provide responses of different quality. Spinning reserve is the most valuable, and therefore generally the most expensive, service because it requires the generator to be on line and synchronized to the grid. Because such generators are online, they can begin responding to a contingency immediately; that is, their governors sense the drop in Interconnection frequency associated with the outage and begin to increase output within a second (Fig. 1). Supplemental reserve, which could include generators that are already online, is less valuable because it does not necessarily provide an *immediate* response to an outage. Replacement reserve is less valuable still because it need not respond fully until 30 or 60 minutes after being called upon. Replacement reserves are used to permit the restoration of the 15-minute contingency reserves so that these faster-acting resources are, once again, able to respond to a new emergency.

REAL-TIME ENERGY

The system operator needs resources that it can use to balance the system in real time (Hirst 2001). This balancing function is typically performed once every several minutes using a computer model that minimizes the cost of meeting electricity demand with the resources then online or that can be started within the next several minutes.

---

*Ancillary services are those functions performed by the equipment and people that generate, control, and transmit electricity in support of the basic services of generating capacity, energy supply, and power delivery (Hirst and Kirby 1998; NERC 2001a).

Typically, suppliers can bid into the real-time market from day ahead through about 30 minutes before the operating hour. The bids into the day-ahead market often include startup and no-load costs, but the bids into the real-time market include only the costs to sell incremental energy or to buy decremental energy. In principle, loads should be permitted to and should be able to participate in this real-time market. In practice, most retail customers will be either unwilling or unable to modify their production processes, comfort levels, or other operations on such a short-notice and frequent basis.

INSTALLED CAPABILITY

The markets discussed so far in this section deal with the short term, from day ahead to actual operations. From a reliability perspective, these markets help system operators maintain security. The long-term equivalent, needed to create adequacy, includes markets for installed generating capability[*] and transmission rights. Traditionally, vertically integrated utilities built or bought the rights to enough generation to meet a loss-of-load probability of not more than one day in ten years. This criterion ensured that the utility had enough generating capacity to meet peak demand with a probability of 99.97%. (One divided by ten times 365 is 0.000274.)

The installed-capability requirements and markets implemented by the three Northeastern ISOs have all experienced problems. A fundamental problem with the requirement is the lack of a tangible product. “Iron in the ground,”which could include a generator that is not able to produce energy (i.e., is unavailable), is of no value during a reliability emergency. A possible product could be the right to convert installed capacity into energy at a predetermined strike price under certain conditions. Instead of recognizing MW of installed (or even unforced) capacity as a reliability product, system operators could require load-serving entities to obtain the rights to energy on demand, a true option. Another possibility is to replace installed-capability requirements with long-term contracts for contingency reserves. Because contingency reserves are well defined products that grant the system operator the right to convert capacity into energy under specific conditions, such long-term contracts might satisfy the ICAP goals in a more efficient manner. Developing demand-side analogs to installed capability will be difficult until the need for and definition of installed capability are clarified.

Many utilities offer their large industrial and commercial customers interruptible rates, roughly equivalent to a long-term call option for contingency reserves. These programs typically offer a discount in the demand charge (expressed in $/kW-month) in exchange for the right to interrupt service to a portion of the customer’s load. These programs are characterized by a rigid structure that specifies months or years in advance the maximum number of times a year the utility can call for interruptions, the minimum amount of advance notice it must

[*]Considerable disagreement exists over the value of an installed-capability requirement in competitive markets; see Hirst and Hadley (1999) and Hobbs, Inon, and Stoft (2001) for perspectives on these issues. See ISO New England (2001) for a discussion of the issues and problems in New England.

provide, the maximum time permitted for each interruption, and the penalty imposed on customers who do not meet their contractual obligation to interrupt demand when called upon to do so. For weather-sensitive loads (e.g., air conditioning and heating), it can be difficult to determine months in advance how much load can reasonably be curtailed. In many cases, these programs were discounts in disguise and were never intended to be used for reliability purposes.

The California electricity crisis of 2000 and 2001 demonstrates well the problems that can occur with these traditional utility interruptible-load programs (California Public Utilities Commission 2001a). Although the three California utilities had been paying industrial and large commercial customers more than $220 million a year for interruption rights, when the California electricity crisis occurred, actual operation of these programs "identified serious problems." In particular, many of the customers participating in Southern California Edison's program did not interrupt their loads as required, resulting in a compliance rate of only 60 to 70% (achieving about 1,200 MW of load reduction instead of the 1,800 MW under contract). In addition, many customers dropped out of the programs as soon as they could, once they realized that the utilities' contractual rights would, under emergency conditions, be exercised.

The three northeastern ISOs have installed-capability requirements and companion markets. Long-term contracts for load interruptions generally qualify as installed capability. PJM's Active Load Management 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 load-serving entities receive installed-capability credits for the load reductions, which reduce their costs of installed generating capacity. Participating loads must be available for up to ten 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 two hours of notification to the load-serving entity by PJM. The baseline is either the customer's load one hour before the event or the customer's hourly load on a comparable day, as determined by the load-serving entity. 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.

Almost 2,000 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, not at all during the summer of 2000, and provided 1800 MW of load relief in 2001 (PJM 2001).

PARTICIPATION OF RETAIL LOADS IN THESE MARKETS

The explanations of these markets suggest that retail loads should be readily able to participate in the day-ahead market for energy, the day-ahead markets for the three reserve services, and the long-term markets for installed capability. Participation in the markets for regulation and real-time energy seem much more problematical for loads because these functions require the ability to modify loads frequently (several times an hour) with only a few minutes advance notice.

Loads that choose to participate in the day-ahead market would bid (either directly or through their load-serving entity) a price-responsive demand curve as shown in Fig. 3. Accepted bids do not limit the customer to consuming electricity in real time as bid. Indeed, changes in weather and other factors (e.g., a new rush order for production of more widgets) might require the customer to consume more or less electricity in real time than was contracted for in the day-ahead market. The customer is free to use however much electricity it wants to in real time. The difference between its actual consumption and its day-ahead schedule (as settled in the day-ahead market) is paid for at the real-time price.

Large municipal water-pumping systems typically have tanks, reservoirs, or lakes to store water for later distribution to consumers (Kueck 2002). These water-storage systems are a natural energy-storage system because they permit the water-treatment system to interrupt pumping operations for up to a few hours at a time. During such nonpumping periods, gravity will ensure sufficient water flow and the appropriate pressure to consumers. The California Department of Water Resources, as an especially important example, has pumping loads of more than 1,500 MW. In aggregate, municipal water pumping accounts for about 3 to 4% of total U.S. electricity consumption (Kueck 2002).

Thus, these municipal systems are wonderful candidates for provision of spinning reserves, both because of their large storage capacity and because of their large size. In aggregate, the nation's water-pumping load is about as large as the nation's spinning-reserve requirement. Recall, however, that current NERC standards prohibit the use of such pumping loads from providing this service!

Permitting these municipal loads to provide (sell) spinning reserve could also improve the efficiency of local water operations. The best way to provide spinning reserve from these loads would be to install adjustable-speed drives on the large pumping motors. The use of such drives, rather than throttling valves, would maintain high levels of pumping efficiency, provide much greater control over pumping operations, and reduce maintenance costs.

In addition to providing spinning reserves, these pumping loads, depending on their locations, might also be able to provide congestion relief and voltage support if they are located in areas with highly-loaded transmission lines and low voltages.

Water-pumping loads are not the only ones that could provide the reserve services. Pumped-storage hydroelectric projects, when in the pumping mode, can provide large amounts of reserves very quickly. For example, a pumped-storage facility in Connecticut has four 250-MW pumping motors. Turning off the pumps could provide up to 1,000 MW of reserves. Some of these systems can switching rapidly from pumping to generation and, therefore, provide almost double the nameplate rating. More generally, any customer facilities or processes with large storage capacity (ranging from residential water heaters to industrial processes that store energy-intensive intermediate product) are good candidates for provision of spinning and supplemental reserves.

Table 2 summarizes the requirements for retail loads to participate in the relevant markets discussed above. Aggregation refers to a minimum size requirement the RTO might impose on all entities connected to the grid. Smaller loads wanting to participate in these markets need to be aggregated so the total exceeds the RTO's minimum size requirement.

Participation in the day-ahead energy market requires meters that can record and store hourly consumption data. Participation in the ancillary-services markets might, however, require meters that can record and store subhourly consumption data, perhaps at the 5-minute interval. The ISO or load aggregator needs this detailed information to determine whether individual customers responded to the ISO's call for the reserve service and did so within the required time. Aggregation of some load reductions, such as the use of switches to turn off water heaters, will likely not require interval meters. If enough such loads are aggregated that the load reduction is "visible" to the system operator in real time and the system operator is permitted to test this resource periodically, meters that are read monthly should suffice.[*#]

Communications include the transfer of information from the RTO to the customer and the transfer of information from the customer to the RTO. The former include acceptance of bids in the various markets, the associated market-clearing prices, and the real-time calls to provide the ancillary services purchased in the day-ahead markets. The latter include bids into the energy and ancillary-services markets and information on actual electricity consumption.

[*]Periodic testing will determine how reliable a resource this load reduction is (in particular, what fraction of the connected load reduction actually responds to activation) and how the magnitude of this resource varies with certain factors such as weather (e.g., water heater interruptions will provide more load relief in the winter than the summer) and time (e.g., water heaters are more likely to be on early in the morning when people first get up).

[#]The load aggregator must still develop an equitable method for sharing the payments from the RTO to the participating customers, recognizing that not all customers will respond every time.

Table 2. Characteristics of load participation in wholesale power markets

	Day-ahead energy	Spinning reserve	Supplemental reserve	Replacement reserve
Aggregation	RTO might require minimum size, say 1 MW, which would require aggregation for all but the larger industrial loads			
Meters	Hourly	Interval meters capable of recording consumption at the 5- or 10-minute level		
Communication	Daily submission of hourly bids to RTO, daily receipt of hourly prices	Daily submission of hourly capacity and energy bids to RTO, RTO must be able to call on winning bidders to reduce loads within required times		
Advance notice	Day ahead	10 or 15 minutes		30 or 60 minutes
Frequency		Customers are free to participate in these markets as they choose; once having chosen on a day-ahead basis to sell reserves during certain hours, they are then committed to providing that service if called upon		
Duration	Not applicable	The load reductions might need to be sustained for as long as an hour (spinning and supplemental reserves) or two hours (replacement reserves)		
Penalties	None	Penalties applied because load committed to make reductions upon RTO call for reliability service (quid pro quo for reservation payment)		
Payments	Day-ahead market price for energy	Day-ahead market clearing prices for capacity plus energy payments for actual load reductions when called upon		
Baseline	None	Because advance notice is so short, baseline is usually consumption during one or a few intervals before the ISO call		

[a]Where generators are permitted to bid in startup and no-load costs, retail customers should be permitted to bid in curtailment-initiation costs to reflect costs they might incur in getting ready to modify their hourly electricity use.

Advance notice refers to the amount of time the customer has to respond to a particular request. Participation in the day-ahead energy market imposes no requirement on the customer. Customers can consume as much or as little electricity in real time as they want; however, any

differences between these actual amounts and those contracted for in the day-ahead market are settled at the real-time energy price rather than the day-ahead price. Customers selling load reductions as ancillary services, on the other hand, *must* comply with the RTOs request to reduce load. Depending on the service, the leadtime available is 10 or 15 minutes for the two contingency reserves or 30 or 60 minutes for replacement reserves. Failure to comply, either in magnitude or in time, will result in a penalty. At a minimum, the RTO will likely withhold the day-ahead capacity payment for the service; in addition, it might impose a penalty related to the reliability risk imposed on the system by the load's failure to comply.

There is no explicit payment for participation in the day-ahead energy market. Rather, the customer pays for the contracted amount at the day-ahead hourly price. Typically, customers selling load reductions into the reserve markets are paid for the capacity they have agreed to provide, a reservation payment in $/MW per hour. In addition, if the load is called upon in real time to provide the reserves, it is also paid for its energy reduction, typically at the higher of its day-ahead energy bid or the current market-clearing price for energy.

Some demand-response programs require definition of a baseline against which the actual load is measured to determine the amount of load reduction. This concept does not apply to the day-ahead energy market. And it is straightforward to apply to the reserve markets because the amount of time between the call for the reserves and their delivery is so short. Typically, consumption during a few 10-minute periods before the call is used to define the baseline. Alternatively, the baseline is defined as the average consumption during the same hour for several previous days of the same type.

Participation in markets for installed capability provide a long-term equivalent to participation in markets for operating reserves. A customer could choose to participate in an installed-capability program (such as PJM's) and receive monthly payments for the capacity it provides. In return for this stream of payments, the customer makes a long-term commitment (i.e., sells an option on interruption rights) to the supplier, permitting the supplier to take the contractual amount of load relief under prespecified conditions. These conditions typically include the number of times a year or month interruptions can be called, the minimum advance notice to be given to the retail customer, the maximum duration of the interruption, and the penalties for failure to comply.

Participating in the day-ahead reserve markets, on the other hand, involves much more modest and short-term obligations. The customer is required to interrupt load when called upon only during those hours on the following day for which its bid was accepted. The penalty for noncompliance is likely to be much more modest also because the commitment time is so much shorter (hours rather than months or years).

The NERC (1993) demand-side management document asks 13 interesting and important questions about the use of loads as reserves. Although the document was prepared

before competitive markets for reserves were developed, they are still applicable to the design and implementation of demand-side reliability programs:

> (1) Is DSM [demand-side management] under operator control?
> (2) Is DSM armed [available] at all times?
> (3) Is DSM under push button (supervisory) control?
> (4) Is a phone call required to activate the DSM?
> (5) Is advance notice required to activate?
> (6) If yes to (5), how much time is required to activate?
> (7) What is the length of time from activation of the DSM until complete load control is achieved?
> (8) Is DSM temperature sensitive?
> (9) Is DSM demand sensitive?
> (10) Can the customer override the activation of DSM?
> (11) Is there a limit on the number of times a day or week DSM can be activated? Is there a limit (practical or contractual) on the duration of the DSM control cycle during a day?
> (12) Does the utility know in real time how much demand reduction can be achieved by DSM activation? How is this information gathered?
> (13) Does the system operator have authority to activate DSM without further approval?

CURRENT ISO DEMAND-SIDE MARKETS

The ISOs began operating special demand-side programs for reliability services during summer 2000. The ISOs refined and expanded their programs (in particular, to include economic demand-side programs) for summer 2001 (Table 3).

These special reliability markets procure the equivalent of contingency-reserve services from loads.[*] The ISOs likely established these separate markets, rather than encouraging retail loads to participate in ancillary-service markets, for a variety of reasons. PJM has no markets for reserves, and the New England markets are fundamentally flawed. More generally, market participants likely felt that the metering and telecommunications requirements for participation in what had historically been generation-only functions, were too onerous. Whether these demand-management pilot programs disappear after the ISOs create or modify their markets to encourage retail-load participation or whether they become permanent is unclear.

[*] I write "equivalent" because the one- or two-hour advance notice in the PJM and New York programs is much longer than the 10-minute notice usually associated with contingency reserves.

Table 3. Comparison of ISO summer 2001 reliability load-reduction programs

	PJM	New York	New England
Payment basis	Actual load reduction		Reservation payment plus payment for load reduction
Payment amounts	Higher of zonal price or $500/MWh		Energy clearing price
Availability of load reductions	9 am to 10 pm		8 am to 10 pm
Minimum capacity per customer	0.1 MW, interval meters required		
Penalties	None		Refund of reservation payment to later of first of month or last successful performance
Baseline	Load during hour before PJM calls for load reduction	Average of hourly loads during highest 5 of the last 10 weekdays	Average of hourly loads during past 10 business days
Dispatch	Maximum emergency generation	Operating Reserve Shortage or Major Emergency	Operating Procedure 4
Advance notice	one hour	two hours	30 minutes, up to 2-hour duration

Source: Biewald and Johnston (2001).

Because these programs are new, their effects have been modest (Goldman, Heffner, and Barbose 2002). For example, only 7 MW of demand reduction signed up for the New England program. The PJM reliability program achieved a maximum reduction of 62 MW one day in August 2001, with an average reduction of 21 MW (PJM 2001). And the New York program achieved an average reduction of 355 MW over 23 event hours (Neenan Associates 2002).

4. INVOLUNTARY LOAD REDUCTIONS

As discussed above, the system operator has many resources it can use to prevent a reliability problem from becoming a catastrophe. However, under certain circumstances, the system operator may be forced to decide between involuntarily interrupting some customers or placing the bulk-power system at great risk. Interrupting some customers can help restore the necessary generation:load balance to maintain system security. On the other hand, operating

the system at or beyond stated reliability limits exposes the system to a possible collapse. Restoring the bulk-power system can take hours or days and can be very expensive.

As a consequence, all electrical systems have various system protection and control systems. If the use of contingency reserves and other routinely available resources is insufficient to meet security requirements, the system operator will manually interrupt some loads and institute a rolling blackout. If even these actions are not enough, additional loads will be automatically interrupted. As spelled out in the NERC (1997) *Planning Standards*, underfrequency load shedding is "required to help protect the security of the generation and interconnected transmission systems during major declining frequency events." These switches, strategically placed throughout the transmission grid, automatically interrupt some customers when the frequency drops too far below its reference value of 60 Hz. In a similar fashion, some loads are interrupted automatically if local voltages drop too much, through the use of undervoltage load shedding.

Although some of these systems are automatic and some require operator action, they share a common feature—the customers that are forced to provide this valuable reliability service by suddenly doing without electricity are uncompensated. Although we pay generators (and a few loads) for contingency reserves, we do not pay loads for providing the ultimate backstop security measure.

The present system is inequitable in that it, once again, treats generators and retail loads differently.[*] Perhaps more important, if system operators paid customers whose loads were involuntarily interrupted, it would likely spark strong interest among customers in the creation of markets for these reliability services. Consider what might happen if the ISO decided to pay, say, $1000/MWh when customers were involuntarily interrupted because of problems on the bulk-power system. Some customers for whom the cost of an interruption is very high might offer to pay $2000/MWh to avoid being called upon in these situations. Other customers for whom the cost of electricity is more important and reliability less important might volunteer to be interrupted at a much lower price, say $500/MWh. These customers offers would then stimulate the creation of active markets in involuntary load reduction. The prices set in these markets would then replace the administrative determinations made by system operators or government regulators (either zero or some arbitrary amount).

California's Optional Binding Mandatory Curtailment Program (California Public Utilities Commission 2001b) contains some of the features proposed here. Customers that agree to reduce their loads by 15% are exempt from stage 3 rotating blackouts. Although this program does not permit customers to buy their way out of involuntary outages, it does permit them to

[*]Some ISO market participants oppose emergency demand-response programs because they see these programs as redundant and an unnecessary expense. After all, they argue, loads can be involuntarily interrupted without compensation because that is how systems have always operated.

avoid such outages by committing to reduce loads during system emergencies. Failure to meet the load-curtailment requirement results in a penalty of $6,000/MWh for all excess energy.

Creating such markets will introduce difficult implementation issues. For example, if an outage last for more than an hour, it could be difficult to determine the baseline consumption for which a customer will be paid. However, bringing retail loads into reliability markets is sufficiently important that I hope regulators, system operators, and other electric-industry participants will consider this idea.

5. CONCLUSIONS

This paper describes current North American reliability rules and practices. Because of the traditional belief that retail customers could not and would not want to manage their loads in response to economic incentives, these rules and practices generally exclude loads from the provision of, and payment for, reliability services. The paper shows that this need not and should not be the case. Expanding reliability functions and markets to include retail loads will improve bulk-power reliability or lower the costs to maintain reliability at current levels. New computing, communications, and control technologies will expand the scope and reduce the cost of using retail loads to provide reliability services.

Achieving this desired end state will require changes, which leads to three recommendations:

- NERC and the ten regional reliability councils should continue to review their reliability rules to remove any technology bias. That is, the standards should specify what is to be done but not how that reliability goal is to be achieved. They should replace the prescriptive elements in their standards with performance standards. And these performance standards should be written so that customer loads are permitted to provide those reliability services they can technically and economically provide.

- ISOs, utilities, other retail service providers, state regulatory commissions, FERC, and others should educate customers on the potential benefits they would derive from participating in the markets for ancillary services, especially for the contingency reserves.

- System operators should pay retail customers whose loads are involuntarily interrupted to preserve bulk-power reliability. Such customers include those subject to underfrequency relays, undervoltage relays, or a rotating blackout.

ACKNOWLEDGMENTS

I thank Richard Cowart, Charles Goldman, Michael Jaske, Brendan Kirby, David Lawrence, David Nevius, and Tracy Terry for their very helpful comments on a draft of this paper.

REFERENCES

B. Biewald and L. Johnston 2001, *The Other Side of Competitive Markets: Developing Effective Load Response in New England's Electricity Market*, Synapse Energy Economics, Cambridge, MA, June 13.

S. Braithwait and A. Faruqui 2001, "The Choice Not to Buy: Energy $avings and Policy Alternatives for Demand Response," *Public Utilities Fortnightly* **139**(6), 48–60, March 15.

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

California Public Utilities Commission 2001b, *Order Instituting Rulemaking into the operation of interruptible load programs…*, Decision 01-04-006, Rulemaking 00-10-002, San Francisco, CA, April 3.

J. D. Chandley 2001, "A Standard Market Design for Regional Transmission Organizations," *The Electricity Journal* **14**(10), 27–53, December.

R. Cowart 2001, *Efficient Reliability: The Critical Role of Demand-Side Resources in Power Systems and Markets*, National Association of Regulatory Utility Commissioners, Washington, DC, June.

C. Goldman, G. Heffner, and G. Barbose 2002, *Customer Load Participation in Wholesale Markets: Summer 2001 Results, Lessons Learned, and "Best Practices"*, draft, Lawrence Berkeley National Laboratory, Berkeley, CA, February.

E. Hirst 2001, *Real-Time Balancing Operations and Markets: Key to Competitive Wholesale Electricity Markets*, Edison Electric Institute, Washington, DC, and Project for Sustainable FERC Energy Policy, Alexandria, VA, April.

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.

E. Hirst and B. Kirby 1998, *Unbundling Generation and Transmission Services for Competitive Electricity Markets: Ancillary Services*, NRRI 98-05, National Regulatory Research Institute, Columbus, OH, January.

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

B. F. Hobbs, J. Inon, and S. E. Stoft 2001, "Installed Capacity Requirements and Price Caps: Oil on the Water, or Fuel on the Fire?" *The Electricity Journal* **14**(6), 23–34, July.

ISO New England, Inc. 2001, *Compliance Filing*, Docket No. EL00-62-015, submitted to the Federal Energy Regulatory Commission, Holyoke, MA, June 4.

J. Kueck 2002, personal communication, Oak Ridge National Laboratory, Oak Ridge, TN, January 3.

R. Lafferty, D. Hunger, J. Ballard, G. Mahrenholz, D. Mead, and D. Bandera 2002, *Demand Responsiveness in Electricity Markets*, Office of Markets, Tariffs, and Rates, U.S. Federal Energy Regulatory Commission, Washington, DC, January 15.

Neenan Associates 2002, *New York ISO Price-Responsive Load Program Evaluation: Executive Summary*, prepared for the New York Independent System Operator, Altamont, NY, January 15.

North American Electric Reliability Council 1993, *Demand-Side Management, The System Operator's Perspective*, Princeton, NJ, December.

North American Electric Reliability Council 1997, *NERC Planning Standards*, Princeton, NJ, September.

North American Electric Reliability Council 2001a, *Reference Document: Interconnected Operations Services*, Interconnected Operations Services Subcommittee, Princeton, NJ, March 28.

North American Electric Reliability Council 2001b, "Policy 1 — Generation Control and Performance," Version 2 of proposed policy change, Princeton, NJ, June 17.

North American Electric Reliability Council 2001c, "Policy 1 — Generation Control and Performance," *NERC Operating Manual*, Princeton, NJ, November 15.

North American Electric Reliability Council 2002, *The NERC Functional Model, Functions and Relationships for Interconnected Systems Operation and Planning*, Princeton, NJ, January 20.

PJM Interconnection LLC 2001, *Report on the 2001–2002 PJM Customer Load Reduction Pilot Program*, Norristown, PA, submitted to the Federal Energy Regulatory Commission, Docket No. ER01-1671-000, December 28.

U.S. Federal Energy Regulatory Commission 1996, *Promoting Wholesale Competition Through Open Access Non-Discriminatory Transmission Services by Public Utilities*; *Recovery of Stranded Costs by Public Utilities and Transmitting Utilities, Final Rule*, Docket Nos. RM95-8-000 and RM94-7-001, Order No. 888, Washington, DC, April 24.

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