New England Demand Response Initiative

TECHNICAL ISSUES RELATED TO RETAIL-LOAD PROVISION OF ANCILLARY SERVICES

Background Issues Discussion

Brendan Kirby and Eric Hirst*

February 10, 2002

1. Introduction

Responsive load is the most underutilized reliability resource available to the power system. It is currently not used at all to supply spinning reserve.† This background issues paper discusses the technical requirements for load, or any resource, to provide contingency reserves to the electric power system. It discusses why some loads may be ideal providers of contingency reserves; why some loads are better at providing spinning reserve than they are at providing any other type of response. It examines the existing reserve resource mix in New England and begins to quantify the benefits *to the power system* of encouraging loads to provide spinning reserve. It discusses the differences between load and generation as spinning reserve resources and ways to accommodate loads while increasing power system reliability. Examples of available load response technologies are provided with a discussion of how they could be used for spinning reserve.

This paper concentrates on spinning reserve, as opposed to all contingency reserves or all ancillary services for three reasons. First, New England may have a market for spinning reserve before it has markets for other ancillary services. Second, loads *are* different than generators. The tempting approach of incrementally adapting ancillary service requirements which were established when generators were the only available resources, will not work. While it is easier for most generators to provide replacement power and non-spinning reserve (the slower response services) than it is to supply spinning reserve (the fastest service) the opposite is true for many loads.‡ Similarly, nonperformance risks and monitoring requirements (and costs) are fundamentally different for loads than they are for generators. Starting with the slower reserve services and incrementally adjusting requirements as experience is gained will block the most attractive responsive loads from ever supplying ancillary services because the loads are not able to sustain the longer response required for the slower services and prices of the slower reserves are much lower than those for spinning reserve. We need to find other ways to develop this resource and build system operator confidence.

---

* Brendan Kirby is a senior researcher at the Department of Energy's Oak Ridge National Laboratory in Oak Ridge, TN. Eric Hirst is a Consultant in Electric-Industry Restructuring in Bellingham, WA.

† Pumped-storage facilities are sometimes used as spinning reserve while in the pumping mode but these are more like generators than loads. They are large individual facilities with full utility instrumentation and control. Most importantly, they are primarily in the energy business.

‡ The limited amount of storage available to most loads limits the response duration.

Responsive load can be as reliable and robust a resource as generation. However, the way it achieves that robustness and reliability is through aggregation of numerous independent loads rather than through the impressed commitment of a few generators. To obtain the full economic and reliability value of responsive load to the overall power system the rules that govern contingency reserves need to be addressed.

2. Reserve Requirements

The electric power system is unique in that aggregate production and consumption must be matched instantaneously and continuously. Several types of controllable reserves are maintained to help the system operator achieve this required generation/load balance. Regulating reserves compensate for the continuous random minute-to-minute fluctuations in load and uncontrolled generation. Frequency-responsive reserves compensate for the frequency deviations. The daily cycling of load is compensated through load following and generator dispatch. Finally, sudden failures of generation and transmission are addressed with three additional reserve products: spinning reserve, supplemental reserve, and replacement reserve (collectively referred to as "contingency reserves").

Conceptually the generation/load balance can be maintained by controlling generation, load, or both. Historically, system operators have tended to control generation almost exclusively. Generators are typically in the business of providing their services to the power system so their business model (whether they are owned by an integrated utility or are independent) accommodates following system operator directives. Communications and control technology also made it easier to monitor and control a few large resources than numerous smaller resources. Consequently the rules governing how the power system is operated were developed at a time when large generators were essentially the only resources available to support system reliability. Rules were prescriptive as to the actions to be taken and the technologies to be used rather than being results oriented (i.e., performance based).

Restructuring has changed the business relationships between generators and the system operator. Technology has advanced to allow loads to be responsive. Energy costs have risen and have become more volatile from hour-to-hour, which provide incentives for loads to respond. Rules established by regulators and technical organizations are being changed to accommodate this new set of circumstances.

2.1 Technical Requirements

While responsive load can theoretically provide almost any service the power system requires (black start may be the only exception), most loads are best suited to provide contingency reserves. Contingency reserves restore the generation/load balance after the sudden unexpected loss of a major generator or transmission line. Power system frequency drops suddenly when generation trips, as shown in Figure 1. In these instances, there is no time for markets to react. In this case frequency sensitive generator governors responded immediately to stop the frequency drop. Spinning and supplemental reserves successfully returned frequency to 60 Hz within ten minutes. Power systems typically keep enough contingency reserves available to compensate for the worst credible event (contingency). This is typically the loss of the largest generator or the largest importing transmission facility. In Texas the simultaneous loss of two nuclear plants is credible (as shown by the event recorded in Figure 1) so the Electric Reliability Council of Texas requires over 2600 MW of contingency reserves. Frequency response, spinning,

supplemental, and replacement reserves operate in a coordinated fashion, as shown in Figure 2.

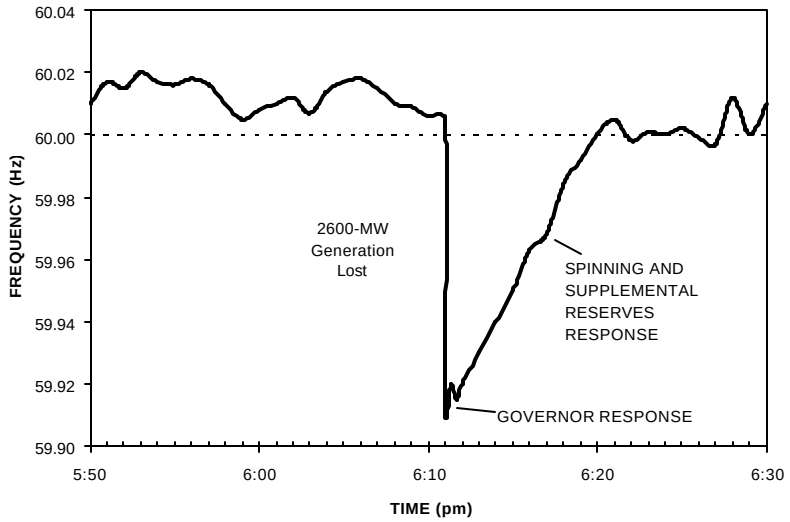

Figure 1 Governor response and contingency reserves successfully restored the generation/load balance after the loss of 2600 MW of generation.

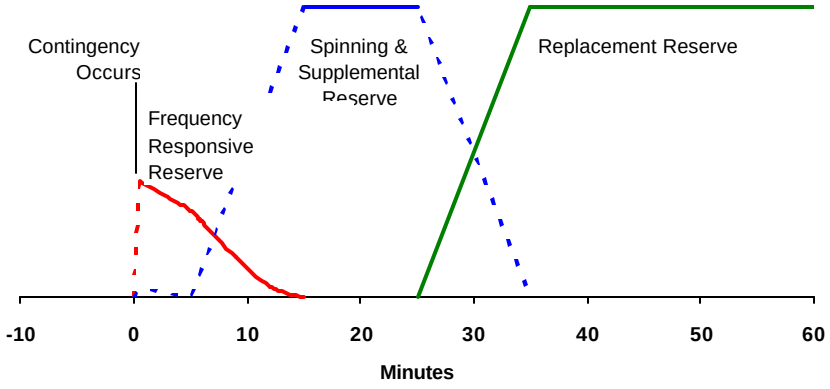

Figure 2 Contingency reserves provide a coordinated response to a sudden loss of supply.

2.2 Regulations and Policies

While the general concepts of system operations and reliability are well established implementation details continue to evolve as the industry is restructured. The Federal

Energy Regulatory Commission (FERC), the North American Electric Reliability Council (NERC), the Northeast Power Coordinating Council (NPCC), and ISO New England (ISO-NE) all have rules and procedures that govern contingency reserve requirements. These rules are not yet consistent among organizations but the trend towards open, technology neutral market based solutions is clear.

2.2.1 FERC

The Federal Energy Regulatory Commission (FERC 2002a), in its notice on Standard Market Design (SMD), shows a clear preference for market-based solutions for energy supply and reliability. They also encourage demand participation on an equal footing with generation. The proposed SMD specifies day-ahead markets for spinning and supplemental reserves, but not for the 30-minute replacement reserve. These markets are to be integrated with the energy market, much as New York does. FERC also proposes operation of real-time markets for ancillary services, again, much as New York proposes in its Real-Time Scheduling system.

2.2.2 NERC

In its most recent operating manual (NERC 2002) NERC has continued its move away from prescriptive requirements for operational practices to relying more on performance standards. 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 (CPS) covers normal operations and the Disturbance Control Standard (DCS) deals with recovery from major generator or transmission outages.

Policy 1 still discusses the resources that control areas will need to meet the performance standards. Each control area is required to have sufficient operating reserves to "account for frequency support, errors in load forecasting, generation loss, transmission unavailability, and regulating requirements". It defines "sufficient operating reserves" as "the capacity required to meet the Control Performance Standard (Section A), Disturbance Control Standard (Section B), and Frequency Response Standard (Section C) of this Policy".[**]

NERC's DCS is a performance measure; it specifies that the control area must recover the generation/load balance within 15 minutes of the start of a contingency. To provide resources to meet the DCS Policy 1 defines contingency reserves as a subset of operating reserves:

> Each CONTROL AREA shall have access to and/or operate CONTINGENCY RESERVES to respond to DISTURBANCES. This CONTINGENCY RESERVE is that part of the OPERATING RESERVES that is available, following loss of resources by the CONTROL AREA, to meet the Disturbance Control Standard (DCS). CONTINGENCY RESERVE may be supplied from generation,

[§] Policy 1 contains five additional standards: Frequency Response and Bias, Time Control, Automatic Generation Control, Inadvertent Interchange, and Surveys.

[**] The frequency response standard will likely evolve into a performance standard similar to CPS and DCS but it is currently still only a specification of how to set the frequency bias.

controllable load resources, or coordinated adjustments to INTERCHANGE SCHEDULES.[††]

Policy 1 goes on to state that each regional reliability council will establish contingency reserve policies covering the minimum reserve requirements, the mix of spinning and supplemental reserves, and "the limitations, *if any*, upon the amount of interruptible load that may be included" (emphasis added). There is a further requirement that each control area or reserve sharing group carry at least enough contingency reserves to cover the most severe single contingency.

There are two important points here. First, the composition of the reserves is not specified. NERC no longer requires spinning reserves to come from generation (though regional councils are not prohibited from setting that requirement).[‡‡] Second, contingency reserves are to be used to meet the DCS standard. That is, they are to respond to contingencies. If they are used to respond to forecast errors, generation or transmission maintenance, or other such problems, they are not available to respond to contingencies. This latter distinction is important to responsive loads because it has a large impact on the response duration. Oddly, as we will discus in greater detail, responsive loads, unlike most generators, care about what the response is to be used for.

There are two other issues relevant to NERC policy. First, meeting the CPS requirements (balancing generation and load under normal conditions on a minute-to-minute basis) uses the regulation ancillary service. 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-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.

Second, frequency response requirements are evolving within NERC policy. Clearly system frequency is important under both normal and contingency conditions. Frequency is the most ubiquitously available measurement of system health (it can be observed at any household 110-volt outlet). The automatic generation control system (AGC) uses regulating resources to precisely control system frequency under normal conditions. Generators larger than 10 MW are required to have active governors that respond to frequency deviations. A frequency control standard similar to CPS and DCS will likely be included in Policy 1 in the future but it is not there yet.

Some responsive loads have the *potential* to be excellent providers of frequency responsive reserve under contingency conditions.[§§] As will be discussed later, they can provider faster and greater response than most generators. The cost of providing that response can also be quite low if the requirement is designed into the load control system.

[††] NERC capitalizes terms in their policies with NERC defined meanings.

[‡‡] The "Terms and Definitions" in the NERC Operating Manual have not yet been updated and spinning reserve is still defined as "unloaded generation that is synchronized and ready to serve additional demand."

[§§] This discussion is of intentional/deliberate frequency control action. Many motor loads and most synchronous generators also have a natural frequency response that aids in maintaining system stability.

However, frequency response is not a function that comes naturally to most loads. While there are many reasons to control loads remotely, and those remote control systems can be readily converted to provide spinning reserve, there are no similar reasons for loads to have frequency response capability built in. Hence there is no pool of frequency responsive loads immediately available with which to run a major test or to start a responsive load program. Load equipment designers would have to see a long-term market for frequency response before they started to design the capability into either the loads or load control systems. In addition to specifying the response details, designers would have to see a stable set of requirements before frequency response is built into equipment designs.[***]

2.2.3 NPCC

NERC's DCS is a performance measure; it specifies what must be accomplished (recovery within 15 minutes) without specifying how that goal must be reached. NPCC has about 6 reportable DCS events per month and only 2 DCS violations in the past 2 years. NPCC requirements are more prescriptive concerning how much reserve is required and what constitute reserves. The reserve definitions are slightly different as well. Table 1 provides the NPCC reserve definitions (NPCC 2002a)

Table 1 NPCC Contingency reserve definitions (emphasis added) (quoted from NPCC 2002a)

Reserve Type	Description
Operating	The sum of the ten-minute and thirty-minute reserves
Ten-minute	The sum of the synchronized and non-synchronized reserve that is fully available in ten minutes
Thirty-Minute	The sum of the synchronized and non-synchronized reserve that can be fully utilized in thirty minutes, excluding capacity assigned to ten-minute reserve
Synchronized	The unused portion of **generating capacity** which is synchronized to the system and ready to pick up load to claimed capacity and capacity which can be made available by curtailing pumping hydro units.
Non-Synchronized	That portion of operating capacity, which is available for synchronizing to the network and that capacity which can be made available by applying load management techniques such as curtailing interruptible loads or implementing voltage reductions.

NPCC's current definition of synchronized reserves is technology specific (restricted to generation with an exception for pumped hydro) rather than being performance based.

[***] What the requirements are (respond in 20 cycles vs. 1 second or respond at 59.96 Hz vs. 59.94 Hz, for example) will likely have little impact on the cost of new equipment, but retrofitting existing equipment to accommodate changing requirements may be prohibitive. Adding frequency response capability after the fact will also be more expensive than designing the capability into the equipment in the first place.

Other than the requirement that synchronized reserve comes from generation the performance requirement is identical to non-synchronized reserve.

The amount of reserves each control area is required to maintain is somewhat performance dependent. Control areas that consistently maintain good DCS performance can reduce the amount of synchronized ten-minute reserve they carry and substitute non-synchronized ten-minute reserves. Good performers are required to have 20% of their ten-minute reserves synchronized while poor performers must have 100% of their ten-minute reserve synchronized. The NPCC Operating Reserve Criteria (NPCC 2002b) specifies how much of each type of reserves must be maintained (Table 2).

Table 2 NPCC Contingency reserve requirements

	10-minute synchronized reserve	10-minute non-synchronized reserve	30-minute reserve
Amount required	25%-100% of first contingency (depending on control area performance)	75%-0% of first contingency (depending on control area performance)	50% of second contingency
Acquired by ISO-NE	600-700 MW	600-700 MW	600-1000 MW
Minimum sustainable time	60 minutes	60 minutes	60 minutes
Maximum restoration time	90-105 minutes	90-105 minutes	4 hours

The NPCC Operating Reserve Criteria and Glossary of Terms recognize the importance of maintaining system frequency but they do not impose a frequency response requirement on contingency reserves. Frequency response is addressed by generators with AGC providing regulation.[†††]

2.2.4 ISO New England

ISO New England has a peak demand over 25,000 MW and typically acquires about 600 to 700 MW each of spinning reserve, supplemental reserve, and replacement reserve. The ISO began acquiring more than 1000 MW of replacement reserves in May 2002. The

[†††] We are not sure why the trend in NPCC and NERC seems to be towards placing the frequency response requirement with the resources providing regulation. Individual frequency response occurs only after frequency has deviated significantly (outside the 0.035Hz governor dead-band). This will only occur as the result of a serious contingency. It seems more appropriate to tie frequency response to the contingency reserves. The logic for making frequency response a regulation requirement may be that the *resources* providing the regulation service (generators under AGC) are more likely to be *capable* of governor response.

extra amount is intended to make explicit the ISO's former implicit day-ahead commitment of resources to meet possible additional reserve requirements. The amounts vary from hour to hour and from month to month; for example, the total amount of reserves acquired during January 2002 ranged from 1270 to 2000 MW, with an average of 1730 MW. The largest single contingency in New England is usually the loss of a nuclear unit (typically operating at 1100 MW) or the loss of the DC tie to Quebec (operating at up to 1500 MW). Reserve response and duration time requirements are probably influenced by the generation resource mix which has few quick-start units and numerous slow-ramping thermal units.

ISO-NE does allow load to provide spinning reserve:

> Ten Minute Spinning Reserve (TMSR) (1) is the Operable Capability of a Generator that is unloaded, is in excess of the quantity required to serve current demand, is able to begin immediately to supply energy to serve demand, is fully available within ten minutes and is able to be sustained for a period equal to the longer of thirty minutes or published NERC or NPCC requirements; (2) is capacity and energy supplied to pump storage generators operating in the pumping mode that can be shut down or whose demand can otherwise be reduced within ten minutes and remain off line or reduced for the longer of thirty minutes or published NERC or NPCC requirements; and (3) is capacity and energy supplied to a Dispatchable Load whose demand can be partially or totally reduced within ten minutes and remain so reduced for the longer of sixty minutes or published NERC or NPCC requirements. (ISO-NE 2002)

Unfortunately responsive load seems to have been tacked on as an afterthought. While there are detailed descriptions of how generator operating limits and ramp rate are to be handled, how lost opportunities are to be compensated, etc. there is no similar detail provided for responsive loads. Further, it is not clear why ISO-NE requires responsive load to maintain spinning reserve response for 60 minutes while it only requires generators to maintain response for 30 minutes. Allowing loads to participate is an important first step in making requirements technology neutral but it is only the first step. The requirements need to be based on genuine power system requirements not simply on the capabilities of the historic supplier. That is not to say that the capabilities and limitations of resources should not be considered. They should. The system operator's objective should be to maximize the benefits to all system users by drawing in as many resources as possible. Service requirements should accommodate and take advantage of as many diverse resources as possible.

3. Responsive Load as a Spinning Reserve Provider

Spinning reserve has traditionally been supplied from generators. It can be thought of as providing insurance. You get paid for having it available whether or not it is called upon to address a system emergency. It should be called upon only in the event of a genuine system emergency such as a lost transmission line or failed generator. Also, it should be called upon infrequently, perhaps a few times per month. It should be restored to service rapidly (within 30 to 60 minutes) so that the power system is protected against the next contingency. Having the option to use load to supply spinning reserve would provide another source of revenue for responsive loads, increase the reliability of the electricity supply by increasing reserves, and decrease all customers energy bills because reserve generation would be freed up to supply energy.

The basic idea of providing spinning reserve (or any contingency reserve) from load is quite simple (Kirby and Kueck 2000):

- Rapidly curtail load in response to a DCS event
- Maintain the curtailment until the next category of reserves (non-spinning followed by 30 minute supplemental) are deployed
- Restore the reserve (and load) to service as rapidly thereafter as possible to be ready for the next contingency

This requires loads with both storage and control capability, a communications system that tells the loads when to respond, and monitoring to assure that the load response was obtained. It also requires a system of rules and tariffs that encourage and accommodate load response. It is important to bear in mind that only a small percentage of loads need respond. ISO NE only requires spinning reserve equal to about 4% of total load.

It is critical to evaluate the *integrated* cost of load response to the load providing the response. Costs will vary from time-to-time depending on the activities the load is engaged in. Most loads must maintain control over their own operations, responding to electricity and ancillary service prices as appropriate at that time.

3.1 Desirable Characteristics

Many different types of loads can potentially supply contingency reserves to the power system. Good candidates share common characteristics:

3.1.1 Storage

Any load that inherently has some storage in its process, or any process to which storage can be readily added, is a good candidate to supply contingency reserves. This will not be direct storage of electricity – 60 Hz power can not be stored directly. The storage will typically be in the form of the productive effort in which the load is normally engaged. Candidates include thermal storage (building heating/cooling, water heating, refrigeration and freezing, etc.), process inventory, compressed air, water pumping, and probably dozens of other such uses.

A common characteristic for many loads is that storage is limited. A building operator may be able to curtail space conditioning for 15-60 minutes, but would be unable to sustain that total curtailment for hours at a time (especially on hot, humid days during which electricity systems tend to peak). Similarly, municipal water pumping systems may have 24 hours of storage available but be unable to curtail for more than an hour in response to power prices or demand reduction requests. The bulk of the stored water must be reserved in case of a large fire or prolonged power outage to address the load's primary mission.

Other load storage capability favors longer outages. A factory with sufficient inventory to forgo production could simply shut down production for a shift or two. There may be higher costs associated with shutting down and restarting production that can only be justified over relatively longer interruption periods. Longer notification times may be required as well so that the work force can be rescheduled. These loads are better suited to respond to the energy market or to provide sustained emergency demand reduction than they are to provide contingency reserves.

3.1.2 Control Capability

To be useful as a contingency reserve the load must have the capability to respond to curtailment requests. The rapid response requirements require automatic response. Loads that already have control systems that can be adapted to respond to spinning reserve commands are best. Control may be secondary control rather than directly cutting power to the load. Simply tripping the circuit breaker that supplies a load can be effective, but may come at a higher cost and potentially cause damage. Adjusting the temperature set point for a thermal load or providing a bypass for a compressor may be as fast and more effective.[‡‡‡]

3.1.3 Notification Requirement

Contingency events are rapid. Immediate response is best. Processes that require no notification time are best suited to providing spinning reserves. Thermal loads, water pumping, air compressing, and other loads with storage inherently built into the process generally do not require advanced notification of curtailment.

Time can be allotted to allow the power system and the load to negotiate participation in the provision of spinning reserves for a specific time. Day-ahead and hour-ahead markets for contingency reserves generally provide this type of notification of resource selection giving the load ample time to arm the control system. But notification of deployment in case of a DCS event must be as rapid as possible.

[‡‡‡] For example, providing a flow bypass on a compressor system that is activated when a curtailment is needed (rather than shutting off power to the compressor motor) can mechanically unload the motor driving the compressor and reduce the electrical load without actually stopping the compressor and risking damage to seals, motors, and other equipment. The details are specific to the specific load but the concept is general.

3.1.4 Response Speed

Faster is better. Once a load with control capability has been notified to deploy in response to a DCS event, the actual response must be accomplished rapidly. Here load response typically exceeds generator response. Many responsive loads (thermal loads for example) provide full response essentially instantaneously. Some responsive loads may require time for valves to operate and processes to go through shutdown procedures. These loads must be able to meet NERC and ISO requirements for full response within 10 minutes (including communications time which might be slower for small loads than it is for large generators).

3.1.5 Restoration

Once a contingency event has been successfully dealt with, the power system needs to have the contingency reserves restored as quickly as possible to protect against the next potential generation or transmission failure. NERC, NPCC, and ISO rules covering reserve restoration are vague and the restoration times are excessively long[§§§] because they were designed to accommodate the only resources available at that time: large thermal generators. These generators have minimum run times and minimum off times to avoid damaging the units from excessive thermal stress. Most responsive loads inherently have much better restoration characteristics. Typically, loads wants to return to service as quickly as possible after the contingency event, and are immediately available in case of another contingency event; reserves are restored immediately.

This last point concerning immediate reserve restoration demonstrates why the reason for the curtailment is important to the responsive loads. A responsive load with limited storage needs to return its processes to their normal operations (building temperature, for example) after a contingency event. Hence the load will be running and potentially available for re-curtailment. The power system needs to be protected in the unlikely event that another contingency occurs quickly. A responsive load might be willing to assume this risk given its low probability of occurrence and given an understanding that it is in everyone's interest to prevent this next contingency event from cascading into a major blackout of a power system that is in a weakened state. The same load might be entirely unwilling to respond to a simple shortage of supply. The load might fear that its limited contingency response capability would be turned into a prolonged energy source.

Unlike generation which typically can run for as long as required after it is deployed at essentially constant cost, costs for responsive load typically start low but rise rapidly after the inherent storage capability is exhausted. A freezer, for example, can curtail at low cost until the temperature rises to the point here the contents start to spoil. Costs rise dramatically after that point. Different loads will have different cost consequences resulting from extended outages.

[§§§] These response duration requirements are excessive in the sense that DCS requires rebalancing the system within 15 minutes and NPCC 30-minute reserves are to be available within 30 minutes. The 10-minute synchronized and non-synchronized reserves should be relieved by 30-minute reserves within 30 minutes.

A solution that might work well for both the power system and the responsive loads would be to price subsequent responses within a limited time frame at progressively higher (perhaps even dramatically higher) prices. Both the power system and the load would be protected from these low-probability/high-consequence events.

3.1.6 Size

Aggregate size is important. The size of the aggregate resource needs to be large enough to be useful. ISO NE requires 600-700 MW each of spinning, supplemental, and replacement reserves. Size of individual resources is different. Both large and small resources have advantages. Large resources are easier to monitor. Small resources behave statistically and potentially have higher reliability as a group (i.e., the failure of any one resource has little impact on the electric system).

3.1.7 Minimal Cost

Both standby and deployment costs need to be considered. Since loads are not primarily in the energy business they are not specifically designed to respond to power system needs. Selecting loads with inherent control and response capabilities reduces the cost of adding spinning reserve capacity.

3.2 Motivation

Power system operators and power market designers should be motivated to encourage responsive load to participate in contingency reserve markets because this will increase reliability and reduce costs to all power system customers. When loads provide contingency reserves, generation is freed up to provide energy. This increases generation supply, which reduces the energy-clearing price for everyone. Similarly, increasing the resource pool for contingency reserves necessarily reduces their costs to the system. The faster response offered by some loads further increases reliability. Finally, encouraging retail loads to provide reserves reduces the market power, in both energy and ancillary service markets, which some generators might otherwise have.

Providing spinning reserve is a better match to the natural capabilities of many loads than responding to hourly energy prices. There may be enough storage to allow one hour of response but not six or ten. However, providing contingency reserves and peak reduction or responding to hourly energy markets are not mutually exclusive. Some responsive loads can respond to multi-hour adjustments and still have additional capacity available for spinning reserve; 2/3 of the response capacity remains available for spinning reserve response even when Carrier's ComfortChoice thermostats are providing peak load reduction, as is explained later. Costs are reduced in this case because much of the communications and control equipment is used for both functions.

Loads are motivated by the compensation they will receive. In 2002, both NY and California experienced hourly prices for 10-minute spinning reserve that were double the price for 10-minute non-synchronous reserve (Table 3).

Table 3 Spinning reserve hourly prices were twice as high as non-synchronized 10 minute reserve prices in both California and New York in 2002.

	NY East $/MW-h	NY West $/MW-h	CA $/MW-h
Spinning Reserve	3.04	2.82	3.89
Non-Synchronized Reserve	1.51	1.37	1.57

3.3 Responsive Loads Are Different Than Generators

There are three major differences between loads and generators (other than the direction of power flow). First, loads are not primarily in the energy business, they have other obligations. Second, individual loads are typically smaller than individual generators; they provide a statistical rather than a deterministic resource. Third, many loads are better matched to naturally respond to fast, short, infrequent events.

3.3.1 Load's Control of Its Own Destiny

While automatic deployment is necessary when selling reserves, it is often important to allow the customer to decide when it will participate and when it will not. Amazingly this is true for 2000 MW industrial plants, for residential customers, and for most loads in between. Just as the price of hourly energy and each of the ancillary services vary, so do customer economics. For many customers there are times when less flexibility exists and the load cannot be interrupted without high costs being incurred. These times are often independent of anything happening on the power system and are therefore unrelated to the price and value of the service. For the right price, a residential customer might be willing to automatically curtail air-conditioning use for 30 or 60 minutes to supply contingency reserves on most days, for example. This same customer would probably be unwilling to curtail use at any price on the evening when s/he was holding a dinner party, however. Similar restrictions might apply for an industrial customer such as a continuous chemical processing plant while it is taking a monthly inventory and needs a stable process. In both cases the customer choice not to participate is unrelated to the bulk-power operations and wholesale electricity markets; neither load is trying to avoid providing the service when it is highest in value. In fact, the chemical plant may intentionally select times for its inventory when the power system is not stressed, such as at night or on weekends. It would do this not because of a concern for the power system but because that may be a time when the chemical process is stable as well due to reduced activity at the chemical plant.[****]

3.3.2 Statistical Response

Fundamentally load is a statistical resource while generation is a deterministic resource. Some loads are large and deterministic while some generators are small and statistical, but as a general rule loads are small, important in aggregate, and behave statistically while generators are large, important individually, and behave deterministically. There

[****] The Paducah Gaseous Diffusion Plant is such a load. It is normally very responsive to spot market energy prices but it becomes completely inelastic during the brief monthly inventories, willing to respond only to a power system catastrophe.

are advantages to both resources, and both should be used. The important thing to note is that there are differences.

Aggregations of small responsive loads can provide greater reliability than fewer numbers of large generators, as illustrated in Figure 3. In this simple example, contingency reserves are being supplied by 6 generators that can each provide 100 MW of response with 95% reliability. There is a 74% chance that all 6 generators will respond to a DCS event and a 97% probability that at least 5 will respond, which implies a nontrivial chance that fewer than 5 will respond. This can be contrasted to the performance from an aggregation of 1200 responsive loads of 50 KW each with only 90% reliability. This aggregation typically delivers 540 MW (as opposed to 600 MW) but never delivers less than 520 MW. As this example illustrates, the aggregate load response is much more predictable.

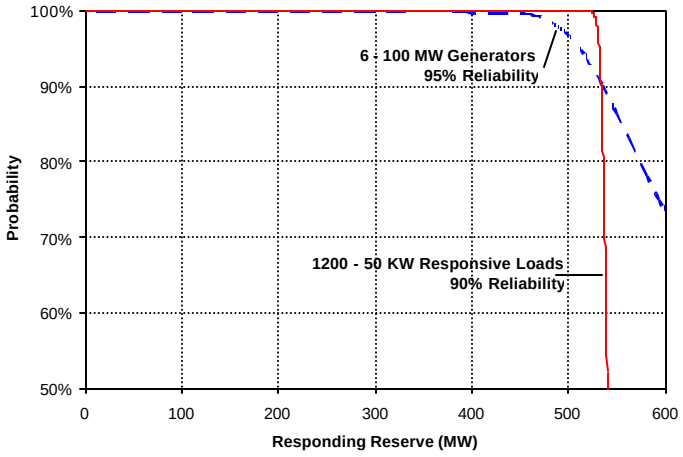

Figure 3 Larger numbers of individually less reliable responsive loads can provide greater aggregate reliability than fewer large generators.

Contingency reserves have historically been provided by large generators that are equipped with SCADA monitoring equipment which telemeters generator output and various other parameters to the system operator every several seconds. Contingency reserve resources are closely monitored for three reasons: 1) to inform the system operator of the availability of reserves before they are needed, 2) to monitor deployment events in real-time so that the system operator can take corrective action in case of a massive reserve failure, and 3) to monitor individual performance so that compensation motivates future performance. Because the same monitoring system provides all three functions, we often fail to distinguish between these functions. For small loads it may be better to look at each function separately. We will examine the requirements in the order of monitoring speed.

Real-Time Event Response Monitoring

Individual generators are typically large enough such that a failure to respond to a call for reserves is a serious event for the system operator. The failure must be addressed

immediately. Other resources must be called upon. Hence it is necessary for the system operator to observe the real-time response of the generator supplying contingency reserves. Even if the failure rate of generator response is very low, the consequences of the failure are great enough to warrant real-time monitoring.

Individual loads are typically small. Failure of any individual load to respond is inconsequential to bulk system reliability. It is only the aggregated response that is important. Real-time observability of individual loads adds little to assuring bulk system reliability, and may not be required. This issue is important because response monitoring, which can be financially supported by large resources, is impractical for small resources. Insistence on real-time monitoring will exclude many small loads from supplying contingency reserves, reduce the reliability reserves available to the system operator, and increase costs. Testing and verification can be used to establish the reliability of the resource. Individual responsive loads can be tested under controlled conditions, both before being allowed to provide contingency response and periodically while in service, to assure that they are capable of providing the required response. Response to actual deployment events can also be verified after-the-fact.

Resource Availability

Some responsive loads can support a monitoring system that informs the system operator of the resource availability. Carrier's ComfortChoice responsive thermostats (discussed later) can report current status but it can take 90 minutes to hear back from all thermostats in a 15,000-unit fleet. Alternatively, forecasting may help provide a highly accurate assessment of available spinning reserve from responsive load. Such forecasts of load response could be based on expected temperature and humidity, day type, and time of day, for example. Reliance on forecasts of aggregated load response is not new. In fact it is aggregated load forecasts that drive the establishment of capacity and reserve requirements. None of the plans, contracts, or commitments for reserve capacity is any more certain than the underlying load forecasts used to design reserve margins. These forecasts are based on expected load response to weather and economic conditions. They are not based on long-term contracts or commitments.

Loads typically cannot guarantee continuous contingency reserve resource availability. The load must be running in order to be available to shut down. So loads can not make a long-term flat commitment to supply reserves, though some loads can participate in day-ahead markets for reserve services. Contingency reserves are typically of highest value when overall load is high and generation is scarce. Thus, the overall statistical load pattern of responsive load combined with unavailability of contingency reserves from generators increases the relative value of responsive load.

Performance Monitoring

Performance monitoring is required. Without some form of performance monitoring it is likely that loads will eventually stop responding since there will not be an incentive to perform maintenance or incur the inconvenience of response. Performance monitoring does not require second-to-second real-time communications, however.

Several options are available. Performance can be monitored at each responsive load and reported back through a slower, cheaper, communications system such as a two-way pager. Alternatively, responsive loads could be tested and certified when they are placed in service and tested periodically and/or randomly thereafter. Alternatively, a small sample of loads can be monitored using interval meters. If the sample is selected properly, results can be scaled up to accurately represent the entire population.

3.4 Communications Requirements

Communications between the power system operator and the responsive load are required for several reasons:

- Resource selection
- Deployment
- Real-time monitoring (to assure the system operator of the availability of the resource and to assist in real-time operations)
- Performance monitoring (to assess performance after-the-fact, set payments, and motivate future performance)

The type of communication, its speed, and the amount of data transferred are different for each alternative. Fortunately, the highest speed requirements are associated with a function (deployment) that has the minimal amount of information and which can be broadcast to all responsive loads (or to a large group) rather than having to be sent to each load specifically.

3.4.1 Resource Selection

Selecting which resources will participate and when is a negotiation process between the power system and each responsive load (and generator). The communication speed required to facilitate this negotiation is fairly slow but the amount of information can be large. In its most general sense this process starts with collecting enough information about potential responsive loads to design the market structure. Once stable markets are in place, this process will typically reduce to the system operator announcing its needs for the day- or hour-ahead, collecting bid information on price and quantity from potential resources, and responding to the selected individuals. Communication requirements are resource specific. Communication time is typically not critical though providing individual notification to tens of thousands of individual loads via individual phone calls would not be practical. The smallest responsive loads will likely opt in and out of a supply program seasonally rather than hourly.

3.4.2 Deployment

The system operator's command to deploy in the event of a contingency is at the opposite end of the spectrum. It must be very rapid. Fortunately it contains very little information, simply an order to respond *now*. The command can be broadcast to the entire resource pool or to an appropriate subgroup. Individual messages are not required. Hence, communications technologies such as radio or pagers that support group notification are better than technologies that exclusively support individual communications such as telephones. A requirement for individual communications may make sense only for the largest resources, which might also state how much response is desired.

3.4.3 Real-Time Monitoring

As discussed above, high-speed real-time monitoring of individual resources is necessary for large resources such as generators and very large loads but is probably not necessary for aggregations of smaller loads that behave statistically.

3.4.4 Performance Monitoring

Monitoring of individual load's performance is necessary to motivate future performance. It is individual communication with a modest amount of information (how quickly the load responded, for example, or availability during all hours that the load was "on call"), but it need not be extremely fast. Such information is needed only within each billing cycle.

3.4.5 Aggregation and Communication

The major objection often voiced to customer supply of ancillary services is that the system operator cannot deal with the large number of individual resources and that the communications requirements would be overwhelming. These are valid concerns but ones that can be addressed. Aggregators can provide a genuinely valuable function here. By handling the communications with a large number of distributed facilities they can present the system operator with a single point of contact for a reasonable amount of capacity, similar to the system operator's interface with large, central generating resources. They can also be an interpreter between the electrical system and customers. The system operator is not interested in learning the details and concerns of each customer. Similarly, customers are in businesses of their own and have neither the time nor the interest in learning all about the power system. The aggregator can bridge this gap, creating a valuable resource in the process.

3.5 Responsive Load Resources

The Peak load in New England is approximately 25,000 MW with the summer peak being slightly higher than the winter. Average electricity demand is closer to 12,000 MW. Electric energy consumption is split 37% (4400 MW) residential, 43% commercial (5200 MW), and 20% industrial (2400 MW), as shown in Table 4. (RDI Powerdat Database)

Table 4 New England has a higher proportion of commercial loads, and a lower proportion of industrial loads than the U.S as a whole or than regions like ECAR.

Region	Residential	Commercial	Industrial
New England	37%	43%	20%
ECAR	33%	28%	39%
Entergy	32%	25%	43%
Continental U.S.	38%	33%	30%

New England is less industrialized than the nation as a whole or than specific other NERC regions and subregions. Interestingly, the average industrial customer size is smaller in New England (100 kW) than in the country as a whole (188 kW) or in ECAR (214 kW). These differences in aggregate and individual resource size may make smaller commercial and residential loads more attractive as responsive reserves in New England

than large industrial loads. Determining which loads might be responsive and how much of that load exists is surprisingly difficult. A fair amount of material is available concerning residential loads, much less for commercial and industrial. The Energy Information Administration (EIA) conducts surveys to provide residential energy use data that can be used to estimate loads in each region of the country (EIA 1999). These data include an estimate of the annual energy use for each household and energy-using device. By estimating the seasonal use of each load (space heating is only done in the winter while water heating is done year round) we developed Table 5 to provide an estimate of the average and peak MW consumption of major potentially responsive major residential load types. 1997 is the last year for which data is available although some preliminary data are now available from the 2001 survey.

Table 5 Potentially responsive New England residential loads provide significant spinning reserve opportunities.

Load	Average Load (MW)	Peak Load (MW)
Space heating	1300	4400
Space cooling	700	2400
Water Heating	1500	1500
Appliances	2700	5500
Refrigerators	500	500

While it is desirable to be able to quantify the responsive load opportunities in each load sector it is not strictly necessary. The amount of reserves required is so small compared with the total load in any sector that it is clear that ample opportunities exist. Within the residential sector alone significant opportunities exist with space cooling, space heating, and water heating. Space conditioning is always attractive in any sector since it inherently includes some storage in the thermal mass of the facility. Similarly, refrigeration is attractive. Air compressing may be attractive, water pumping certainly is. Any function that has some short-term storage and is relatively easy to control is a potential resource. Clearly, the amount of load that becomes available to respond will depend on the available payments and the rules for eligibility and response.

Further research to identify specific spinning reserve opportunities is warranted but progress towards developing responsive load as a reliability resource need not wait on detailed results to determine if the concept is viable.

3.6 Costs

Though markets will eventually determine prices for spinning reserve and the other ancillary services in New England it makes sense to briefly examine the costs of generation providing spinning reserve to see if load supply makes any sense. There are several cost components involved when generators supply spinning reserve: capital cost, operating cost (for both standing by and during deployment), and lost opportunity cost (Hirst and Kirby 1997). Contingency reserves are required continuously, even at the time of system peak load. The amount of capacity needed exceeds the amount required to serve the highest load. While there is excess capacity from idle generation most hours,

capacity beyond the highest energy requirement must be built to supply contingency reserves. Once energy and ancillary services markets reach equilibrium ancillary service market prices will have to recover the capital cost of the capacity associated with providing reserves at the time of system peak. The cost for adding contingency reserve response capability to an energy management system can be very low ($10-$100/kW). This cost is especially favorable when compared to cost of new generation ($500/kW).

4. Responsive Load Examples

We provide two detailed examples of existing loads and load control technologies that demonstrate both the advantages and limitations of responsive load supplying spinning reserve. The first example utilizes a large aggregation of residential and small commercial loads. The second involves very large water pumping loads. Both examples appear to be viable. Though these are specific products or implementations and the details are necessarily specific to the examples, the concepts can be generalized to other products using similar technologies. The purpose of this discussion is to demonstrate the possibility of replicating these specific examples to other end-uses and technologies.

4.1 Control of Residential and Small Commercial Thermostats

Carrier Corporation designed the programmable ComfortChoice thermostat and associated communications infrastructure to provide emergency peak reduction for utilities (Kirby 2003). Skytel two-way pagers are used to transmit a curtailment order to the thermostat and to receive back acknowledgment and monitoring information. The customer gets the advantage of a fully programmable thermostat that is web accessible for programming, remote control and monitoring. The thermostat can be retrofitted to any residential or small commercial installation that uses a wall thermostat. Detailed discussions with Carrier revealed that the technology is fast enough[††††] to provide spinning reserve and, we believe, provides ample monitoring capability (Kolb et al 2002) (Carrier 2002). Preliminary analysis of curtailment test results from 2002 reveal that three times as much spinning reserve is available as peak reduction (LIPA 2002).[‡‡‡‡]

The basic system configuration of Carrier's ComfortChoice controllable thermostat is shown in Figure 4. The system operator interfaces with the resource through a web-based system provided by Silicon Energy. It is easy to use and customize to the needs of individual utilities,[§§§§] which reduces the burden placed on the system operator. One or more pager signals are generated and transferred to the SkyTel pager network. Commands go via satellite to pager towers where they are broadcast to the thermostats. The thermostats take immediate action or adjust their schedules for future action, depending on what the system operator ordered. The thermostats log the order and respond via pager letting the utility monitor the response to the event. The thermostats also collect data every minute on temperature, set point, and power consumption (hourly duty cycle) as shown in Table 6. They retain this information as hourly averages and report it back to the utility. The thermostat itself holds seven days of hourly data.

For a summer load curtailment the system operator might send a command at 9:00 in the morning directing all thermostats to move their set points up 4 degrees starting at 14:00

[††††] The load typically receives the system operator command and fully responds in under a minute, ten times faster than the ten minutes allowed for generators to fully respond to spinning reserve commands.
[‡‡‡‡] That is, a system that can provide about 17 MW of demand reduction can supply about 50 MW of spinning reserve. This is because a demand reduction duration of several hours (which potentially could happen several days in a row) is too long to completely curtail the load while a spinning reserve curtailment is short enough (and infrequent enough) to allow complete curtailment.
[§§§§] The term "utility" is used here to refer to whatever type organization is performing the function of controlling the bulk power system.

and ending at 18:00. But the system operator could send a command directing all thermostats to completely curtail immediately. The command would be received *and acted upon* by all loads, providing full response within about 90 seconds. This is far faster than generator response which typically requires a ten minute ramp time.

Thermostats can be addressed individually, in groups, or in total. This has the important advantage of providing both flexibility and speed. System operator commands that are addressed to the entire resource are implemented through a single page that all thermostats receive. Similarly, fifteen subgroups can be addressed if response is required in a specific area to alleviate a transmission constraint. Thermostats can be addressed individually as well. This is useful for monitoring the performance of the system (each thermostat is checked weekly for a "heartbeat").

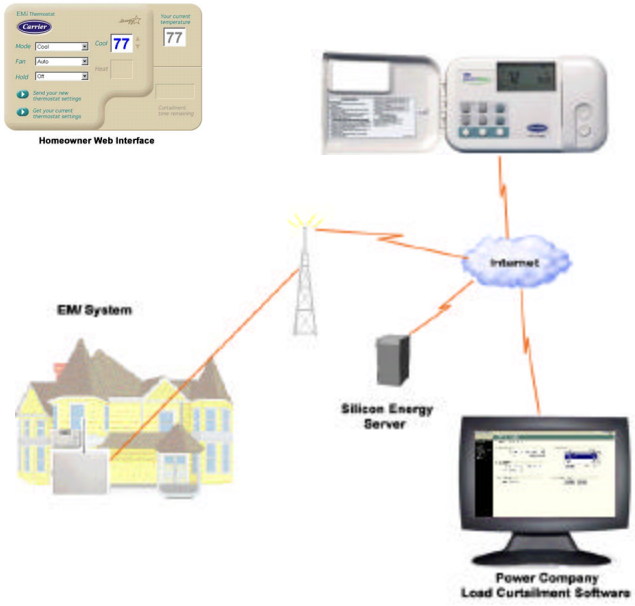

Figure 4 Carrier ComfortChoice thermostats appear to be ideal for supplying spinning reserves.

The customer receives advantages as well. This is a fully programmable, remotely accessible thermostat with all of the associated energy savings and convenience benefits. A web-based remote interface is provided for customer interaction. Customers can also override curtailment events (the system operator can block overrides if they wish but this is not typically done). This feature appears to be important to gain customer acceptance and it probably *increased* the reliability benefit, as will be discussed later.

Two-way paging communication enables the utility to monitor load performance, both during response events and under normal conditions. Response from each thermostat is

staggered over a user-defined interval (where the user is the utility and not the load) to avoid overwhelming the paging system. It typically requires 90 minutes for 15,000 thermostats to respond. Thus the system provides for performance monitoring, but not in the four second intervals typical for large generators.

Table 6 Carrier ComfortChoice thermostats provide significant monitoring capability

Hourly Data
of minutes of compressor/heater operation
of starts
Average temperature
Hour end temperature trend
Event Data
Accurate signal receipt and control action time stamp

Communications are more reliable from the system operator to the thermostat than from the thermostat to the system operator. The pager tower has a 500-watt transmitter while the pager's transmitter is only 1-watt. The thermostat makes four attempts to report back if the pager tower fails to receive any of its signals. The thermostat continues to take control actions and respond to new commands even if return communications are lost. Hence the system is more reliable than would be indicated by the list of "failed" units generated by the "heartbeat" report. Experience to date has found 4%-5% of the thermostats fail to report back.

4.1.1 Experience to Date

A significant number of responsive ComfortChoice thermostats have been installed in the U.S. as shown in Table 7. Long Island Power Authority (LIPA) collected information on the heating and cooling equipment being controlled when they installed 17,000 ComfortChoice thermostats for the LIPA*edge* program (LIPA 2002). They also directly measured the power consumption of a subset of these loads to estimate the actual load of the aggregation. LIPA determined that the average residential air conditioning unit being controlled consumes 3.84 kW while the average small commercial unit consumes 6.38 kW.

LIPA tested the actual performance of the system to reduce energy demand during peak hours on three days during the summer of 2002: July 3, July 30, and August 14, from 2:00pm to 6:00pm. They obtained 15.8, 16.1, and 16.3 MW of demand reduction from 15,943, 17,051, and 17,474 thermostats respectively.[*****] LIPA found that responsive thermostats provide on average roughly 1 kW of demand reduction for each thermostat. LIPA utilizes a 50% duty cycle when curtailing load which limits the demand reduction. We estimated the potential spinning reserve benefit by assuming that the thermostats could be curtailed completely for the shorter duration of a contingency event and

[*****] The installation program was continuing throughout this time.

assuming a conservative 3 kW per thermostat (78% duty cycle).[†††††] This indicates that spinning reserve can be provided even while peak reduction is being provided (though at a somewhat reduced amount), a surprising synergy.

Table 7 ComfortChoice thermostats already control a significant amount of load for peak reduction. Their benefit from providing spinning reserve could be even greater.

Utility	Thermostats	Estimated Peak Reduction MW	Estimated Spinning Reserve MW
LIPA	17000	17	51
Con Edison	10000	10	30
SCE	5000	5	15
SDG&E	5000	5	15

The 2002 program cost for LIPA was $515 per residential customer and $545 per commercial customer. Customer incentives include a $25 one-time payment to residential customers and a $50 one-time payment to small commercial customers. The customer also receives the internet accessible programmable thermostat. Additionally, some customers were likely motivated to participate in order to help reduce power prices and alleviate the summer power crisis on Long Island.

4.1.2 Manual Override

Manual override is another synergistic concept that works well for both power system reliability and small load performance. Energy management is not the primary concern of most loads, especially small loads. This is one of the basic differences between loads and generators. Loads often find it impossible to make firm, long-term curtailment commitments because there is some chance that external events (external to the power system) will prevent them from reducing power consumption when requested. Even if a customer is able to respond 99% of the time the other 1% of the time may be perceived to be of such high importance that the load is unwilling to participate in a curtailment program. This can be true for residential as well as commercial customers.

Day-ahead and hour-ahead hourly markets reduce or eliminate this problem for large loads and generators. But the transaction burden of constantly interacting with energy and ancillary service markets is likely too great for many small loads. Many will prefer to establish a standing offer for response that they are able to honor the vast majority of the time.

Manual override provides an alternative with benefits for both the power system and the customer. With a manual override feature the load curtailment occurs but the individual

[†††††] Initial analysis indicates that temperature rise will not limit total curtailment for spinning reserve response.

customer has the option to override the curtailment. The advantage to the power system is that this increases the load participation and likely reduces the required compensation. The advantage to the customer is that it can opt out of a particular curtailment if the inconvenience or cost for the specific event is unusually high. Many peak reduction programs now include this feature and it appears to be successful. Most importantly, the increase in participation outweighs the number of customers overriding the curtailment.

The natural fear from the power system side is that many customers will always opt out. This is not as large a problem as one might think. Opting out requires the customer to notice that the curtailment is happening and decide that the inconvenience is too great. The customer must take specific action for each event. (DCS events occur roughly once a month.) Customers that chronically opt out could also be dropped from the program.

Manual override is less of a problem when spinning reserve is being supplied than when the peak load is being reduced because the event duration is shorter. Figure 5 shows the override experience of LIPA's 17,000 Comfort Choice thermostats for a peak reduction curtailment on the afternoon of August 14, 2002.[‡‡‡‡‡] (LIPA 2002) Overrides during the first hour were modest.

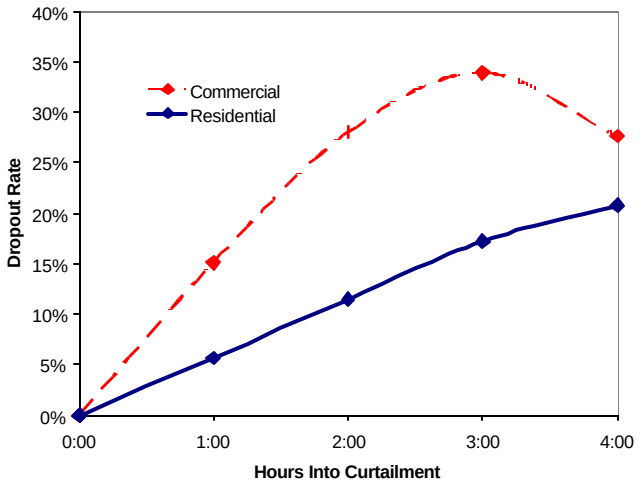

Figure 5 Manual override is not a problem during the spinning reserve time frame.

Carrier's ComfortChoice responsive thermostats offer the additional option of distinguishing between events that the customer can override and events that the customer cannot. This could allow the power system to provide the customer with the ability to opt out of longer demand reduction events while blocking the override during shorter contingency events. The thermostat provides a message on the LCD display indicating "Critical Situation" to let the customer know why they are unable to override the curtailment.

[‡‡‡‡‡] Override data was taken hourly. Though we show overrides starting immediately it is likely that there was some time (perhaps 10 or 15 minutes) before the first overrides started, providing better spinning reserve response.

It is interesting to note that more commercial customers override and they do it more quickly than residential customers. One reason is probably that many residential customers are not home at 2:00 pm when the curtailment events were started. Another reason may be that commercial establishments fear the loss of business if their customers become uncomfortable. This difference in override rate needs to be considered when evaluating response effectiveness, along with the individual load size and load shape.

4.1.3 Responsive Thermostat Summary

Carrier ComfortChoice responsive thermostats are being installed in large numbers for emergency peak reduction. Preliminary analysis indicates that they could be excellent providers of spinning reserve. Roughly three times the load reduction capacity is available for contingency events as is available for peak reduction; two thirds of the capacity is still available to supply spinning reserve when the loads are already curtailed for peak reduction. Adding spinning reserve capability to other load control schemes greatly increases the benefits at little or no additional cost.

Response is likely faster and more effective than generation response as well. Commands are received and response typically completed in 90 seconds as compared with 10 minutes for generation.

There are two areas where this technology does not currently meet strict interpretations of spinning reserve requirements: monitoring speed and frequency response. Frequency response capability could be added relatively cheaply, for perhaps $1-10/device. All of the signaling and expensive control equipment is already there. But this capability will not be added unless manufacturers see that there is a real demand for such capability and that response specifications are established and stable. Retrofitting already installed devices may be prohibitively expensive.

Real-time SCADA monitoring (in 4 second intervals) is the only requirement that is fundamentally difficult to meet. To avoid overwhelming the paging network responses from individual units must be staggered when the entire system is being polled. It can take 90 minutes for 15,000 units to respond. But unlike large generators that can completely fail to respond due to an equipment problem at the generator it is unlikely that 15,000 individual air conditioners or heaters will fail simultaneously. The communications backbone (satellite, cell towers, computer network, etc.) could fail but that can be monitored separately, perhaps at the SCADA rate required for large generators. It is also important to perform an engineering analysis of the communications system to assure that it will not fail specifically when the power system is deploying reserves. Assure that there is sufficient backup power for the communications systems and that reserve deployment messages have priority if the paging system experiences heavy traffic, for example. Establishing monitoring requirements that are appropriate for an aggregation of small resources is needed.

4.2 Control of Large Pumping Loads

At the other end of the spectrum from residential thermostats, large water pumping loads, such as those of the California Department of Water Resources (CDWR), can provide a substantial amount of spinning reserve (Kirby and Kueck 2003). This sophisticated participant in energy and ancillary service markets is now studying providing spinning reserve from its large pumping loads. While this example is specific to California, and we are not aware of any identical loads in New England, the concepts apply to any large load that has some response flexibility and limited storage. Nationally domestic water pumping and irrigation account for about 4% of the electric load. Most of these systems have limited storage. Many other large industrial loads have limited storage as well.

For the CDWR, water delivery is its number one priority, and any plans or modifications must carefully consider this priority. The efficiency and reliability of the water delivery cannot be compromised. However, CDWR's participation in the ancillary services market has provided an important revenue source in the past, and the CDWR is always interested in exploring new ideas. CDWR-owned generation facilities greatly reduce the cost of water pumping. They minimize the amount of power that must be purchased. Generation is scheduled to optimize its value in the energy markets rather than to directly supply the pumping loads.

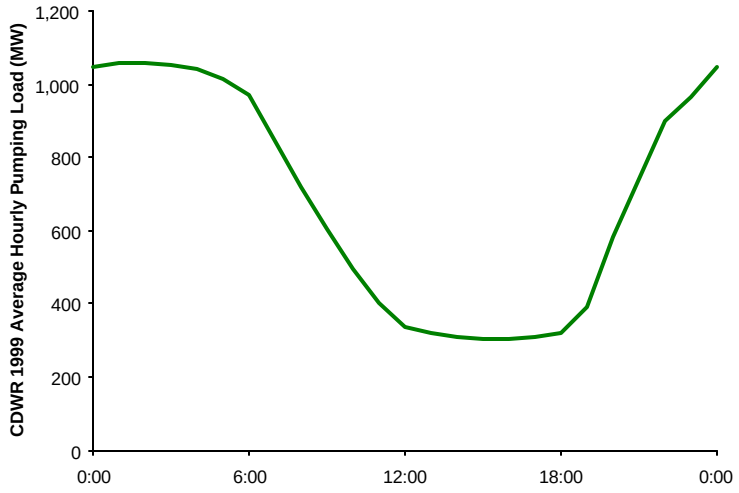

Figure 6 CDWR is a sophisticated manager of its large pumping load, minimizing costs by interacting with the real-time energy market.

The CDWR operates 18 pumping plants, three pumping-generating plants, five hydroelectric power-generating plants, and a coal fired generating plant. In an average year, CDWR generates 5 billion kilowatt-hours and uses 6 billion kilowatt-hours of energy. CDWR generation participated in the Ancillary Service markets (regulation services as well as contingency reserves) when the price makes it worthwhile. To the extent possible, CDWR generates on peak and pumps off peak, as shown in Figure 6. There is also a Remedial Action Scheme where CDWR drops 100's of megawatts in the

event of a system emergency, demonstrating CDWR's commitment to power system security.

Optimizing power purchases to minimize energy costs would appear to limit opportunities for selling contingency reserves. Contingency reserve prices tend to track energy prices and are highest in the middle of the day as shown in Figure 7. Still, CDWR has sufficient potentially responsive load to supply over half of California's spinning reserve requirements. CDWR can supply about a third of the annual spinning reserve requirement even if they decline to sell when the spinning reserve price is below $1.50/MW-hh. CDWR has *more* spinning reserve capacity than California requires for thousands of hours a year. The higher price (over double on average) commanded by spinning reserve provides a strong incentive for selling spinning rather than non-spinning reserve.

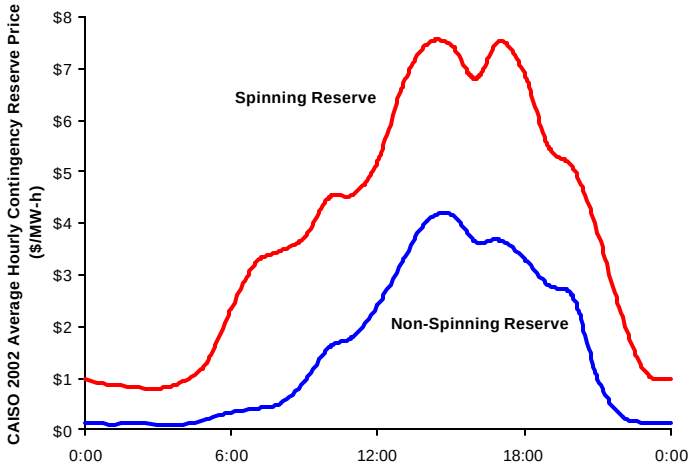

Figure 7 California average 2002 hourly contingency reserve prices peak in the middle of the day, as expected

4.2.1 Pumping Resource

CDWR's pumping facilities are located along some 660 miles of aqueduct (Figure 8). The aqueduct is composed of river channels, reservoirs, canals, pipelines and tunnels. There is little storage along much of the aqueduct and pumps in some locations operate as "strings" to keep the flow even. Turning off one pump at one plant would cause a flow imbalance in the string. A pump must be turned off or on at each plant. There may be sufficient water storage, however, to accommodate the sale of spinning reserves (~30 minutes).

Pump motors, in general, are designed for multiple starts per day though there are limitations.[§§§§§] A rule of thumb for starting an individual motor more than once per day

[§§§§§] Interruptions themselves do not particularly stress the motors but the subsequent restarts do.

has been two starts with a 30-minute cool down in between , and then a one-hour cool down period after any subsequent shutdown. Complex equipment and procedures are required for some of the largest motors. The large Edmonston pumps, 80,000 Hp, use motor generator sets to assist in starting; these motors are not started across the line at rated voltage and frequency. The motor generator sets provide a form of soft start where the motor is started at a reduced frequency.

Figure 8 The geographic diversity of the CDWR pumping system makes it a valuable spinning reserve resource capable of responding even if there are congestion constraints.

4.2.2 Pumping Load Summary

CDWR pumping stations are very large loads. The 54 individual pumps range in size from 350 hp (0.27 MW) to 80,000 hp (63 MW) with an average size of 31,000 hp (24 MW). Providing spinning reserve from large loads is conceptually simpler than supplying reserve from aggregations of small loads. CDWR pumping loads already have communications, monitoring, and control equipment similar to that deployed for generators. Information is telemetered to the CDWR control center, which forwards it to the CAISO control center.

Being large, expensive pieces of equipment, much study and engineering is required before CDWR is willing to risk such resources in providing reliability services to the power system. Capital projects (such as adding soft-start to some pumps) may be required.

CDWR pumping loads have similar advantages and limitations, as do small thermal loads. They can respond quickly, more quickly than the 10-minute ramp time required by

generators. Deployment duration and frequency are both concerns for this large load as they are for aggregations of small thermal loads.

CA ISO system operators and market designers should be eager to work with CDWR. There is a concern, for example, that the CA ISO might not use load-based contingency reserves exclusively to respond to contingencies (the sudden, unexpected loss of a generator or transmission line) or that they might not restore the reserves quickly. It will be important to word contracts carefully so that spinning reserve is called for only in bona fide contingency situations and that it is restored to service quickly. Fortunately, such provisions are in the best reliability interest of both the CA ISO and CDWR.

5. Markets for Contingency Reserves

Since ISO New England began operating real-time markets for energy and ancillary services in May 1999, it has experienced problems with its markets for the reserve services. Complications in the design of the ISO's day-ahead unit-commitment and its 5-minute security-constrained dispatch prevent it from notifying the winning bidders in its ancillary-services markets beforehand. As a consequence, generators do not know whether they were "selected" to provide operating reserves until after the fact. In addition, the ISO might, during a major outage, call on units that were not selected to provide reserves and therefore do not get paid for providing the service.

In August 1999, ISO New England (1999) filed emergency market revisions with FERC. The ISO noted that its first three months of operation had led it to

> conclude that four of the [ISO] markets, ten-minute nonspinning reserve, 30-minute operating reserve, operable capability, and installed capability are fundamentally flawed. They do not require delivery of any physical product, and there is no difference in the costs or risks incurred by those participants who receive payments in the market and those who do not. As a result the only economically rational bid in the market is a bid of zero (to ensure selection in the hope there is any positive price) or a bid that is an attempt to set the clearing price.

In response to the ISO's request, FERC (1999) permitted the ISO to cap the prices of operating reserves at the current hour's energy price. This authority, however, is limited to special circumstance. FERC noted that "… bid caps in the operating reserve markets are limited to periods of capacity deficiency [OP4] or system emergency [OP8] when the ISO is required to choose all bids regardless of how high the price might be. … until this flaw is remedied by an alternative market design, the risk of arbitrarily high prices will remain."

The prices paid by ISO New England for reserves likely have little meaning because of flaws in the ISO's reserve markets. During the past three years, prices have been consistently below $2/MW-hr.[******] Between January 2000 and November 2002, the price of spinning reserve averaged $1.16, the price of supplemental reserve averaged $2.11, and the price of replacement reserve averaged $0.80/MW-hr.

New England will implement a new, improved market design in March 2003, based on the design now operating in PJM. This new market system, however, will not include PJM's two-part market for spinning reserve (see discussion below). ISO New England may have no operating markets for any of the contingency reserves until mid- or late-2003.

[******] MW-hr refers to a megawatt of ancillary service provided for one hour, different from a MWh of energy.

The New York ISO operates an integrated set of markets for energy, real-power ancillary services, and congestion management (Kranz, Pike, and Hirst 2002). Because of the severity of transmission constraints in New York, especially in New York City and Long Island, New York's reserve markets have three zones.

Prices in the New York ISO ancillary-service markets, which do not contain the flaws that the New England markets have, might be a more reasonable indicator of what prices should be in a well-functioning market. New York, like New England, acquires roughly 600 MW of each of the three reserve services each hour. For the 2-year period from July 2000 through June 2002, the prices of spinning, supplemental, and replacement reserve in New York averaged 2.66,[††††††] 1.86, and $0.99 /MW-hr. This ordering of prices is consistent with the value of each service, with spinning reserve the most valuable and replacement reserve the least valuable. (The New England prices, on average, did not follow this order.)

Until December 2002, PJM had no markets for contingency reserves. Any generator committed for service by PJM is guaranteed recovery of the costs associated with unit startup and noload costs. To the extent these costs are not recovered from energy markets during each day, PJM pays these units the difference between their operating costs and revenues for the day. These uplift costs were collected from PJM customers through an operating-reserve payment, although the nexus between these costs and reserves is ambiguous.

Beginning on December 1, 2002, PJM (2002) began operating a two-tier market for spinning reserve. Tier 1 consists of units online, following economic dispatch, and able to ramp up in response to a contingency. These units receive no upfront reservation payment but do receive an extra $50 to $100/MWh for energy produced during a DCS event. Tier 2 consists of additional generating capacity synchronized to the grid, including fast-start combustion turbines that are not generating but are operating as synchronous condensers. These units are paid a reservation charge, based on a real-time market-clearing price but receive no extra energy payment during a reserve pickup.[‡‡‡‡‡‡] PJM does not operate markets for supplemental reserve or for replacement reserve. FERC (2002b) approved the PJM market, noting, however, that it "does not contain all the attributes contemplated by the Commission in the SMD NOPR, and the PJM proposal is different from the spinning reserve markets in New York and New England."

FERC's (2002a) proposed SMD specifies day-ahead markets for spinning and supplemental reserves, but not for the 30-minute replacement reserve. These markets are to be integrated with the energy market, much as New York does. This integration implies that the market-clearing price will reflect both the availability bids of the resource

[††††††] The price of spinning reserve in New York may be slightly higher because this number does not include the opportunity-cost payments the ISO makes to generators that are dispatched below their economic point to provide spinning reserve.

[‡‡‡‡‡‡] It is hard to understand why a competitive market would be designed to pay resources providing the identical service different amounts, and in different ways, based solely on the cost to the resource of providing the service.

plus the location-specific opportunity cost of the resource. FERC also proposes operation of real-time markets for ancillary services, much as New York proposes in its Real-Time Scheduling system. These real-time markets would differ from the day-ahead markets in that potential suppliers would not be permitted to submit availability bids. In other words, the prices for each reserve service in real time would be a function only of the real-time energy-related opportunity costs. FERC is clear that it wants these ancillary-service markets to be open to demand-side resources as well as generators.

6. Issues, Concerns, and Questions

A number of issues remain. These generally require a combination of technical, regulatory, and business solutions. They are briefly discussed here.

6.1 Aggregation

Small loads providing contingency reserves are only of interest if they are aggregated together to obtain a resource that is large enough to have some impact on the power system. The exact power level that is of interest to the system operator is not well defined but it probably lies somewhere between 1.0 and 50 MW. The aggregation function could be performed by any one of several entities. Clearly, system operators themselves cannot be burdened with interacting with thousands of individual loads. But the system operator's organization may want to have direct control of the communications and control infrastructure. Alternatively, a load serving entity or a third party aggregator (i.e., a curtailment service provider) could supply the aggregation service. A large retail chain, such as WalMart might perform that function for the stores it owns. Whoever performs the function must supply rapid communications between the system operator and individual loads without overwhelming the system operator with excessive detail or slowing down the process.

6.2 Forecasting Response

Unlike conventional generation, the availability of responsive load may not be flat or schedulable. There is reason to believe that it can be accurately forecast, however. Forecasting reserve availability from responsive loads is similar to conventional load forecasting. It is both easier and more difficult to forecast responsive loads compared to forecasting total load. It is easier because the responsive loads are more uniform (thousands of air conditioners and water heaters, for example, but not the full diversity of appliances and other residential loads). All of the individuals in an aggregation of air conditioners, for example, are driven by the same weather, time, and day-of-the-week factors. While it is hoped that there will be many types and sizes of responsive loads, there will still be less diversity in this group compared to total load. Forecasting responsive load is also easier because it does not have to be as accurate. A 10% error in a conventional load forecast is a problem because the error may be larger than the entire reserve.[§§§§§§] A 10% error in a load response forecast is not as large a problem because this is less than 10% of the entire reserve.[*******] Also, with continuous performance monitoring, responsive load forecasts will get better with time. Finally, responsive load forecast errors are likely to trend in a beneficial direction. The actual reserve will tend to be greater than the forecast at times when the overall load is also higher than expected and greater amounts of reserve are required (hotter than expected summer afternoons, for

[§§§§§§] The average absolute hour ahead forecast error for the CAISO load was 1.3% in 2002. The average day ahead forecast error was 1.9%.

[*******] A 10% forecast error for a 25,000 MW load is a 2,500 MW problem for the ISO. If responsive load is providing ½ of the 700 MW spinning reserve requirement and the forecast is in error by 10% this results in a 70 MW problem for the ISO. Not a trivial amount but much less than the problem created by the overall load forecast error. The system operator could easily call for 110% response from responsive load if a 10% error was found to be typical.

example). Similarly, when available responsive load is less than forecast it is likely that overall load is also lower than expected and reserve requirements are reduced (cooler than expected summer afternoons, for example). The accuracy and/or importance of these speculations are not yet known but could be investigated as responsive loads are monitored over time.

6.3 Real-Time Monitoring Requirements for Large Generators and Small Loads

When contingency reserves are obtained from large generators it is necessary for the system operator to monitor them in real-time. If a large generator fails to respond the control area will be seriously deficient, will likely fail to meet the disturbance control standard, and will suffer lower reliability. The system operator needs to detect such a failure early enough to be able to direct other generators to respond. If a large generator is 95% responsive it still creates a serious reserve deficiency for the system operator during every 20^{th} event. In the case of generators, therefore, the system operator requires real-time SCADA monitoring.

Small loads do not pose the same problem for the system operator. With tens of thousands of physically independent devices responding, the response can be measured statistically. If each device responds 90% of the time, the aggregation will always provide almost exactly 90% of the available capacity. Real-time-SCADA monitoring is not required in this instance. Instead, the system operator needs to request 10% (in this example) more response than is needed in order to achieve a 100% response rate. Continuous monitoring of common-mode-failure points, such as communications towers, may be desirable.

6.4 Performance Monitoring

Performance monitoring is required. Without some form of performance monitoring it is likely that loads will eventually stop responding since there will be no incentive to perform maintenance or incur the inconvenience of response. Performance monitoring does not require second-to-second real-time communications, however.

Several options are available. Performance can be monitored at each responsive load and reported back through a slower, cheaper, communications system such as a two-way pager. Alternatively, responsive loads could be tested and certified when they are placed in service and tested periodically or randomly thereafter.

6.5 Frequency Response

NERC policies and industry requirements concerning contingency reserves and frequency response are in a state of flux. Clearly, the power system requires frequency responsive reserves. System frequency deviates whenever there is a mismatch between generation and load. A severe transmission or generation contingency will result in a large generation/load mismatch. The consequent shift in frequency is observable anywhere in the power system. Communications between the system operator and the generator or responsive load are not needed. The system frequency change can be detected locally. This is fortunate because response must be very rapid to be useful –

within cycles and seconds rather than minutes. This is the immediate response that prevents the power system from collapsing.[†††††††] Loads can be an excellent frequency responsive resource, in some ways better than generators.

Generators are required to have frequency responsive governors, typically with a 0.036 Hz dead band and a 5% droop. This means that generator controls ignore frequency deviations between and 59.964 and 60.036 Hz. For frequency deviations beyond that, the generator governor is to call for an increase (or decrease) of 100% of rated output for a 3 Hz (5%) shift in frequency. Several factors are important here. First, a 3 Hz shift in system frequency is unthinkable. Total collapse of the system is probably unstoppable if system frequency reaches 57 or 63 Hz. Second, the governor response speed requirements are not clear. The CAISO requires governors to detect frequency within 1 second while NERC uses 20 cycles in some references and one minute in others. Third, generators are only required to provide full spinning reserve response in ten minutes. So while the governor may respond quickly, the generator's full response will lag.

Loads are better and worse at providing fequency response: better because they can provide full response and provide it essentially instantaneously, limited only by the response speed of the detection relay: worse because they do not, individually, provide continuous and reversible response. Having different individual loads respond at different frequencies can simulate continuous droop response.

The real concern is that the rules governing which resources (spinning reserve resources, generators under AGC, or some other designation) must supply frequency response are not clear. Loads typically do not have a reason to respond to system frequency deviations in the governor range. So while loads could be a valuable frequency response resource that capability will not be included in load controllers until there are clear requirements and specifications.

6.6 Deployment Frequency

Typically, the number of responsive load deployments is restricted through explicit numerical limits (interruptions per day, season, or year). It might be better to tie interruptions to specific physical events such as DCS events which typically occur only 10-20 times per year. This provides the power system with the response it needs when contingencies occur, and also provides the load with protection from excessive use for economic or other reasons. A load that wants to respond to economic signals can bid into the energy markets.

6.7 Capacity and Deployment Payments

All resources (generators and loads) have both investment costs and operating costs associated with providing contingency reserve response. There are both monetary costs and investments in time and effort required to be ready to respond. The resource owner

[†††††††] The "natural" frequency response of synchronous generators and motor loads which couples the energy stored in the inertia of rotating equipment to the power system also helps in the very short stability time frame.

has to have sufficient confidence that there will be an adequate market for response to justify the investments. Payment could be for actual response, for hourly reservations when the resource elects to participate, or for seasonal participation. Obligations to respond could be coupled with hourly or seasonal reservation payments. The most likely alternative, consistent with FERC's SMD, is for loads to participate in an RTO's day-ahead market for spinning reserve.

Payment for actual response can work if response is required frequently enough and if prices are consistent enough. This form of payment is less effective if actual response is required infrequently, or if prices remain low for long intervals and only spike occasionally. Black start capability (which is not a service responsive load could sell, but is a good example of an infrequently needed service) would not be paid on an as-used basis, for example. Energy, on the other hand, could be paid on an as-used basis.

Some responsive loads have the inherent capability that they could redeploy almost instantaneously if another contingency happens quickly. The inconvenience and cost to the load rises, however, if it is still recovering from the first contingency. It might benefit both the load and the power system to establish a higher (perhaps significantly higher) price for response to a subsequent event if it occurs within 60 minutes (for example). The power system would be protected from unlikely but high consequence events and the load would be protected from having its contingency response capability turned into supplemental energy. This would represent an improvement for power system reliability over today's practice of taking one to four hours to fully restore reserves.

6.8 Deployment Duration

Generators are typically indifferent to deployment duration.[‡‡‡‡‡‡‡] Lengthening the response requirement from 30 minutes to two or six hours has little impact. A conservative system operator might reasonably request a longer deployment limit for spinning reserve in case it is needed at some time. Loads typically have different deployment constraints. They can provide a large response rapidly but the duration is limited by the available storage. Costs (or damage and inconvenience) often rise dramatically as the outage is extended. On the other hand, loads, unlike fast-start generators, are usually available immediately for re-deployment if a subsequent emergency occurs. It is to the power system's advantage to carefully determine spinning reserve duration requirements in order to draw as many resources into the pool as possible. A deployment interval closer to the 30 minutes necessary to bring replacement reserves on line might be more appropriate than the NPCC's current requirement of 60 minutes.

[‡‡‡‡‡‡‡] Hydro units can be duration limited by water inventory and thermal units can be constrained by emissions limits.

7. Additional Research

Responsive load shows promise of increasing system reliability and reducing costs. Immediate implementation should be actively promoted. Still, there is much that is not known. Research can help increase the value and reduce the cost. Research can also help increase system operator confidence in the resource. A few examples of additional research are provided below.

7.1 Resource Size and Cost

Research is needed to determine what the size of the overall resource is and to identify loads that are particularly attractive for supplying contingency reserves. Response speed, aggregate size, location, and deployment cost (capital and operating) will be important factors.

7.2 Reliability Based On Statistical Behavior

Research is needed to verify that small loads behave statistically and that the aggregate reliability does not require real-time monitoring by the ISO. Tens of thousands of Carrier Comfort Choice thermostats at LIPA, Consolidated Edison, Southern California Gas and Electric, and San Diego Gas and Electric can provide data for this analysis. More importantly, the analysis will be verifying the concept, not just the specific technology being tested. Since the response capability is already installed research to extend the capability from peak shaving to provision of spinning reserve can be conducted relatively quickly.

7.3 Forecasting Responsive Load

Research is needed to develop forecasts of aggregate responsive load behavior. More reliable the forecasting reduces the need for second-to-second monitoring.

7.4 Developing and Testing Market Rules

Research is needed to determine a set of industry structures and performance-based market rules that efficiently utilizes and fairly incorporates responsive loads in contingency reserve markets. Given that one or few entities may be required to capture economies of scale in the aggregation of responsive loads, the degree of centralization in the acquisition of such responsive loads (and associated regulatory oversight) needs to be considered along side market-based approaches.

8. Conclusions

Responsive load has the potential to provide increased power system reliability and reduce costs for all users. The same responsive load can often provide a larger MW amount of spinning reserve response than peak load reduction; three times the amount in one example cited here. This can reduce the cost of contingency reserves and also free up a greater amount of generation to serve load, which would reduce the market-clearing price of energy as well. Advances in communications and control technology now make it possible for even small loads to provide spinning reserve. Responsive load can be as reliable and robust a resource as generation, but the way it achieves that robustness and reliability is through aggregation of numerous independent individuals rather than through the impressed commitment of a few. Making use of this resource requires reexamining the basic contingency reserve requirements and making the rules genuinely technology neutral.

References

Energy Information Administration 1999, *1997 Residential Energy Consumption Survey Public Use Data Files*, U.S. Department of Energy, Washington D.C.

E. Hirst and B. Kirby 1997, *Ancillary-Service Details: Operating Reserves*, ORNL/CON-452, Oak Ridge National Laboratory, Oak Ridge TN, November.

ISO New England, Inc. 1999, *Request for Expedited Approval of Revisions to NEPOOL Market Rules 1 and 10*, Docket No. ER99-4002-000, submitted to the Federal Energy Regulatory Commission, Holyoke, MA, August 5.

ISO New England 2002, *New England Power Pool FERC Electric Rate Schedule No. 6 Market Rules & Procedures*, Holyoke, MA, 27 February

B. Kirby 2003, *Spinning Reserve From Responsive Loads: LIPA*edge */ Carrier ComfortChoice Responsive Thermostats – Initial Results DRAFT*, ORNL/TM-2003/19,Oak Ridge National Laboratory, Oak Ridge, TN, February

B. Kirby and J. Kueck 2003, *Spinning Reserve from CDWR Pumping Load: Initial Results DRAFT*, Oak Ridge National Laboratory, Oak Ridge, TN, February

B. Kirby and J. Kueck 2000, *How Buildings Can Prosper By Interacting With Restructured Electricity Markets*, ACEEE, August

L. Kolb, R. Archacki, and P. Pierret 2002, private discussions with Carrier personnel.

B. Kranz, R. Pike, and E. Hirst 2002, *Integrated Electricity Markets in New York: Day-Ahead and Real-Time Markets for Energy, Ancillary Services, and Transmission*, New York Independent System Operator, Schenectady, NY, November.

Long Island Power Authority 2002, *LIPA*edge, Slide presentation made to the NY ISO Price-Responsive Load Working Group, 21 November

North American Electric Reliability Council 2002, *NERC Operating Manual*, Princeton, NJ, 21 November

Northeast Power Coordinating Council 2002a, *NPCC Glossary of Terms*, Document A-7, New York, NY, November 14.

Northeast Power Coordinating Council 2002b, *Operating Reserve Criteria*, Document A-06, New York, NY, November 14.

PJM Interconnection 2002, "Spinning Reserve Market," Docket No. ER02-2519-000, submitted to the Federal Energy Regulatory Commission, Norristown, PA, August 29.

U.S. Federal Energy Regulatory Commission 1999, *ISO New England, Inc., Order Accepting for Filing Revisions to Operable Capability Market Rules,* Docket No. ER99-4002-000, Washington, DC, September 30.

U.S. Federal Energy Regulatory Commission 2002a, *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 2002b, *Order Accepting Spinning Reserve Market*, Docket No. ER02-2519-000, Washington, DC, October 31.