TECHNICAL AND MARKET ISSUES FOR OPERATING RESERVES

Eric Hirst and Brendan Kirby
Consulting in Electric-Industry Restructuring
Oak Ridge, Tennessee

October 19, 1998

INTRODUCTION

Ancillary services matter! That's the key lesson to learn from the problems faced by the California Independent System Operator (ISO) during the summer of 1998. And California will not be alone. Unless ISO New England changes its market rules, it, too, will face serious problems in the acquisition, pricing, and costs of ancillary services.

These problems arose (and will continue to occur) because the electricity industry as a whole does not yet recognize the technical and market importance of ancillary services. These services are, like transmission, essential for both reliability and commercial functions. Because most of these services are provided by the same pieces of equipment (generators) that produce energy, the energy and ancillary services markets are tightly coupled. Problems in one market will, unavoidably, cause problems in other markets. Poorly designed rules can amplify the problems from one market to the next.

This paper uses operating reserves as an example to illustrate these issues, problems, and possible solutions. Electricity is the ultimate "real-time" product, with its production, transportation, and consumption occurring within a fraction of a second. Because electricity moves at nearly the speed of light and cannot be readily stored, bulk-power systems must maintain a continuous and near-instantaneous balance between production and consumption. Operating reserves are the ancillary services that maintain this balance when a major generator or transmission line unexpectedly fails.

DEFINITION

The utility industry understands well the technical requirements for operating reserves because the service has been an important part of operating power systems for decades (Hirst and Kirby 1997 and 1998). Nevertheless, FERC and NERC described the service differently. NERC's (1997) Policy 1 (Generation Control and Performance) specified the criteria that govern the amount and use of operating reserves:

> Each CONTROL AREA shall operate its MW power resources to provide for a level of OPERATING RESERVE sufficient to account for such factors as errors in

> forecasting, generation and transmission equipment unavailability, number and size of generating units, system equipment forced outage rates, maintenance schedules, regulating requirements, and regional and system load diversity. Following loss of resources or load, a CONTROL AREA shall take appropriate steps to reduce its AREA CONTROL ERROR to meet the Disturbance Control Standard. It shall take prompt steps to protect itself against the next contingency.

FERC, on the other hand, defined operating reserves in Order No. 888 as "extra generation available to serve load in case there is an unplanned event such as loss of generation." FERC defined spinning reserve as the service "provided by generating units that are on-line and loaded at less than maximum output. They are available to serve load *immediately* in an unexpected contingency, such as an unplanned outage of a generating unit." And FERC defined supplemental reserve as "generating capacity that can be used to respond to contingency situations. Supplemental reserve, however, is not available instantaneously, but rather within a short period (usually ten minutes). Supplemental operating reserve is provided by generating units that are on-line but unloaded, by quick-start generation, and by customer-interrupted load, i.e., curtailing load by negotiated agreement with a customer to correct an imbalance between generation and load rather than increasing generation output."

FERC's definition is narrower than NERC's, addressing only the unexpected failure of generation or transmission. NERC allows these reserves to be used to cover generation/load imbalances caused by such factors as load-forecasting errors, generation and transmission equipment maintenance schedules, and load diversity (commercial and forecasting issues) as well as the immediate problem caused by a generation or transmission contingency (reliability issues). But FERC restricts the service to addressing only the reliability concerns. NERC's definition pre-dates restructuring efforts to separate commercial and reliability functions. It is in the process of defining *contingency* reserves as the reliability subset of *operating* reserves. Thus, NERC's definition of contingency reserves may be equivalent to FERC's definition of operating reserves.

Operating reserves typically include two components, spinning and supplemental reserves. NERC's official definitions (taken from the *NERC Operating Manual*) are: *Operating Reserve*—That capability above firm system demand required to provide for regulation, load forecasting error, equipment forced and scheduled outages and local area protection. It consists of spinning and non-spinning reserve. *Spinning Reserve*—Unloaded generation, which is synchronized and ready to serve additional demand. *Non-spinning Reserve*—That operating reserve not connected to the system but capable of serving demand

within a specified time, or interruptible load that can be removed from the system in a specified time. At least half of the operating reserves must be spinning.[*]

The functions to be performed by spinning reserve are not clearly articulated. Requiring reserves to be synchronized, online, and/or to respond immediately says nothing about why these attributes are required. Neither the requirement that spinning reserve begin responding immediately nor that it be frequency responsive has been quantified. This was not a problem in the vertically integrated industry because the characteristics of typical generators were known. Adequate system performance was obtained by requiring frequency responsive governors on the generators supplying spinning reserves. Because the generators were owned by the same entity responsible for control-area security, the generation portfolio would respond as desired. That assurance will not exist in the restructured industry.

NERC's Interconnected Operations Services (IOS) Implementation Task Force is grappling with these issues. The Task Force may recognize that contingency reserves have two *characteristics*: immediate frequency response (a few seconds through 10 minutes) and slower sustained response (10 minutes through 30 minutes). Definitions that recognize these characteristics may help focus the service on system requirements instead of on the physical characteristics of generators.

Some regions (e.g., New York and New England) require additional reserves that must be fully available within 30 minutes. The California ISO requires replacement reserve to be fully available within 60 minutes and then be maintained for two hours. These additional reserves are used to replace the operating reserves to protect against a second contingency. (The Western Systems Coordinating Council requires that operating reserves be restored within 60 minutes after the occurrence of a disturbance.) It is not clear why some regions require replacement reserves and others do not. The NERC requirements deal only with the 10-minute response. However, NERC's (1998) proposed Policy 10 defines secondary reserves as "capacity (including load interruption) that is capable of replacing a transmission customer's scheduled energy supply in the event that it is completely or partially reduced." This secondary reserve serves a purely commercial function.

NERC REQUIREMENTS

Until recently, NERC's standards for "disturbance conditions" consisted of two elements. The B1 Standard required that area control error (ACE) return to zero within ten

[*]In practice, many utilities (e.g., in the Midwest, Florida, and Texas) belong to reserve-sharing agreements. In most of these arrangements, it can take up to a few minutes to notify and activate reserves in neighboring control areas, which violates the "immediate" requirement of spinning reserves.

minutes following the start of the disturbance.* The B2 Standard required that ACE start to return to zero within one minute following the start of the disturbance. The 10-minute full-response requirement associated with operating reserves derived from the B1 standard, while the need for spinning reserve derived from the B2 standard.

As part of its effort to convert from voluntary to mandatory compliance with its reliability criteria, NERC (1997) created a new Disturbance Control Standard that replaces the B1 and B2 standards with a single quantitative measure of control-area performance. Control areas must calculate a recovery factor (R) for all disturbances within the range of 80% to 100% of the control area's most severe single contingency. Control areas are expected to meet the new standard 100% of the time for these reportable disturbances. The recovery factor is defined as:

$$R_i = 100\% \times [MW_{LOSS} - Max(0, ACE_A - ACE_M)]/MW_{LOSS} \ ,$$

where MW_{LOSS} is the MW size of the disturbance as measured at the beginning of the loss,

ACE_A is the predisturbance ACE if $ACE_A < 0$ and 0 if $ACE_A \geq 0$, and

ACE_M is the maximum algebraic value of ACE measured within ten minutes following the disturbance.

To meet the new DCS, the recovery factor must be greater than or equal to 100% for every outage between 80 and 100% of the single largest contingency. Thus, the DCS provides no provision for some excellent recoveries (i.e., $R_i > 100\%$) to offset some poor recoveries (i.e., $R_i < 100\%$). Control areas are required to restore their ACE for every outage within the specified range. On the other hand, outages of less than 80% or more than 100% of the largest expected contingency are not counted at all; control-area operators need not rush to respond to such outages according to the DCS. The need for immediate response to restore frequency is no longer captured in NERC requirements because the DCS imposes no recovery requirement less than 10 minutes, unlike the old B2 standard.

We were unable to locate any data and analysis to support NERC's new DCS or to explain the relationship between reliability and the standard. Discussions with NERC staff identified only one document related to the DCS. NERC (1996) published a set of "frequently asked questions" on the new Control Performance Standard (CPS) and DCS. This document contains only one page on the DCS. More important, the questions and answers provide no documentation on the technical basis for this standard. The basis for the 10-minute recovery

*ACE is the instantaneous difference between actual and scheduled interchange between the control area and the rest of the interconnection, adjusted for any difference between actual and scheduled interconnection frequency (usually 60 Hz).

period (rather than, say, 8 or 12 minutes) is not explained, nor is there an explanation for why 100% compliance with DCS is critical; indeed, as discussed below, current control-area performance falls short of this requirement. The document provides no explanation for the 80 to 100% range for reportable disturbances; as shown below, recovery times are often longer than 10 minutes for smaller disturbances. Finally, the DCS lacks support for the requirement that additional reserves must be provided if a control area fails to achieve 100% compliance.[*]

[*]NERC's new CPS1 and 2 provide an interesting counterexample. NERC and EPRI sponsored a research project that provides a solid basis for these new standards. The project included collection and analysis of detailed data on frequency deviations in all three interconnections and it applied statistical tools to develop CPS1 and 2 (Jaleeli and VanSlyck 1997).

REGIONAL REQUIREMENTS

The minimum operating-reserve requirements differ from region to region (IOS Working Group 1997).* In the East Central Area Reliability Coordination Agreement (ECAR), the spinning and supplemental reserve requirements are both 3% of the daily peak load. In the mid-Atlantic region, the spinning reserve must be the greater of 700 MW or the capacity of the largest unit on line; its supplemental reserve requirement is 1700 MW. In Florida, the spinning reserve must equal 25% of the largest unit online, and the supplemental reserve must equal 75% of the largest unit. In the Electric Reliability Council of Texas, the responsive-reserve requirement is set at 2300 MW. In the other regions, the requirement is based on either the largest generating unit online or the single most severe contingency (either a large generator or a critical transmission facility).

Only limited technical support for the minimum operating-reserve requirements set by the various regions is publicly available from the reliability councils, power pools, and individual utilities (Hirst and Kirby 1997). The lack of data and analysis for the differing regional requirements parallels the lack of technical support for the DCS itself.

The requirement for minimum levels of operating reserves could be based on either probabilistic or deterministic calculations. In most regions, the requirement is deterministically calculated based on the N – 1 criterion. Thus, these requirements are independent of the reliability performance of the generating units in the particular region. The Western Systems Coordinating Council (1997) is the only region with an operating-reserve requirement that recognizes differences among generator types. It requires reserves equal to 5% of the load supplied by hydroelectric resources plus 7% of the load supplied by thermal generation. However, no region incorporates the reliability of *individual* units in its determination of the minimum operating-reserve requirement.

METRICS

Metrics are needed to determine how much reserves are required and to measure a resource's performance in supplying those reserves (NERC 1998). The operating-reserves performance measure is "the extent to which [it] meets the DCS requirement." The performance measures for suppliers of operating reserves include:

- Certification tests to demonstrate the capability of the resource and establish its eligibility to participate in the reserve market,

*We are not sure whether reliability requirements determine minimum reserve levels or vice versa. In practice, regional reliability councils set minimum reserve requirements, and these reserves and their use determine reliability levels.

- The extent to which the resource delivers the required capacity within the required time,

- The extent to which the resource maintains delivery of the required capacity for the required duration, and

- The extent to which the resource controls its return to the pre-contingency schedule.

Metrics are needed for the capability, the rate of response, and the degree of control. Capability refers to the real-power capability (in MW) of the reserves, and responsiveness means the rate (in MW/minute) at which the supplier can respond to the request for reserves. Control refers to the ability to maintain real-power delivery within a specified range of the requested reserve.

RESERVE FUNCTIONS

The discussion of operating reserves so far has emphasized their use to protect against major generation and transmission outages, a form of insurance intended to help maintain bulk-power reliability. In practice, utilities have historically used operating reserves for additional reliability and commercial purposes (Vice 1998).

On occasion, generating units cannot operate at their rated capacity level. As examples, if the coal pile is wet, if the temperature of the inlet cooling water is high, or if some of the plant's equipment is not functioning properly, the generator may not achieve full output. When a unit is returned to service after an extended outage (e.g., for scheduled maintenance), it may not come back online as scheduled. In addition, such a unit is more likely to experience forced outages during the first few days of operation than after it has been operating for some time. Finally, a utility may be using economy (nonfirm) purchases to supply energy to its customers. Because these purchases can be quickly recalled by the seller to meet its reliability requirements, the purchasing utility may need to carry extra operating reserves to back up these nonfirm purchases.[*] Thus, utilities, from time to time, may need to carry more operating reserves than the amount specified by the regional reliability council.

In many cases, the NERC and regional requirements for operating reserves include services beyond those associated with protection against generator and transmission outages. Specifically, these reserves are often used to adjust for load-forecasting error and sometimes include the generation assigned to regulation. The largest uncertainty about tomorrow's load

[*]Reserves to back up 100% of the economy purchases are generally not needed. Such a conservative approach is warranted only when it is likely that all these sales will be recalled at the same time.

is the weather. If, in the summer, it is hotter than expected, air-conditioning loads will be higher. (For southern utilities, a one-degree-Fahrenheit difference between actual and expected summer afternoon temperatures might change the load by 1% or more.) If, in the winter, clouds suddenly move into the service area, lighting loads will increase.

In addition to the disparate reliability and commercial functions of the use of reserves, the timing characteristics of their uses are completely different as well. Operating reserves that cover unit outages must respond within seconds or minutes and without warning. The resources providing this service will be relieved of this function within 30 to 60 minutes and returned to their reserve status. Responding to load-forecast errors occurs over hours rather than minutes, involves advance warning as the error grows from hour to hour, and requires the use of the resource for several hours at a time.

GENERATOR OUTAGES

We contacted several utilities, power pools, and regional reliability councils to obtain data on outages and the time to recover from those outages (Hirst and Kirby 1997). Specifically, data for each control area on the following variables at the time of each outage was requested:

- the number of generating units online
- the number of units providing spinning and supplemental reserves
- the amount of generating capacity committed to operating reserves
- the amount of generation lost because of the outage
- the amount of operating reserves lost because of the outage
- the current hourly and daily peak loads
- the current spot price of electricity (or system lambda)
- the time to restore ACE and frequency to their predisturbance values

Ultimately, we obtained basic data on outages and recovery time for only three control areas. These data show substantial differences in the number of, extent of, and recovery from generator outages (Fig. 1). Relative to the minimum operating-reserve requirement, the outages in ERCOT and New York are much smaller than those for American Electric Power (AEP). AEP not only had larger outages, it also had much longer recovery times. For example, ERCOT averaged 8 outages per year for which the recovery time exceeded the 10-minute standard, New York averaged less than 1 such outage, while AEP averaged 32 such outages. AEP's longer recovery times may be related to the fact that a large percentage of its generation comes from big units: 66% of its system capacity is from units with capacities of 500 MW or more, compared with 50% for ERCOT and 26% for the New York Power Pool.

Analysis of these data shows that the size of an outage is a statistically significant determinant of outage-recovery time. These data also show that many factors beyond outage

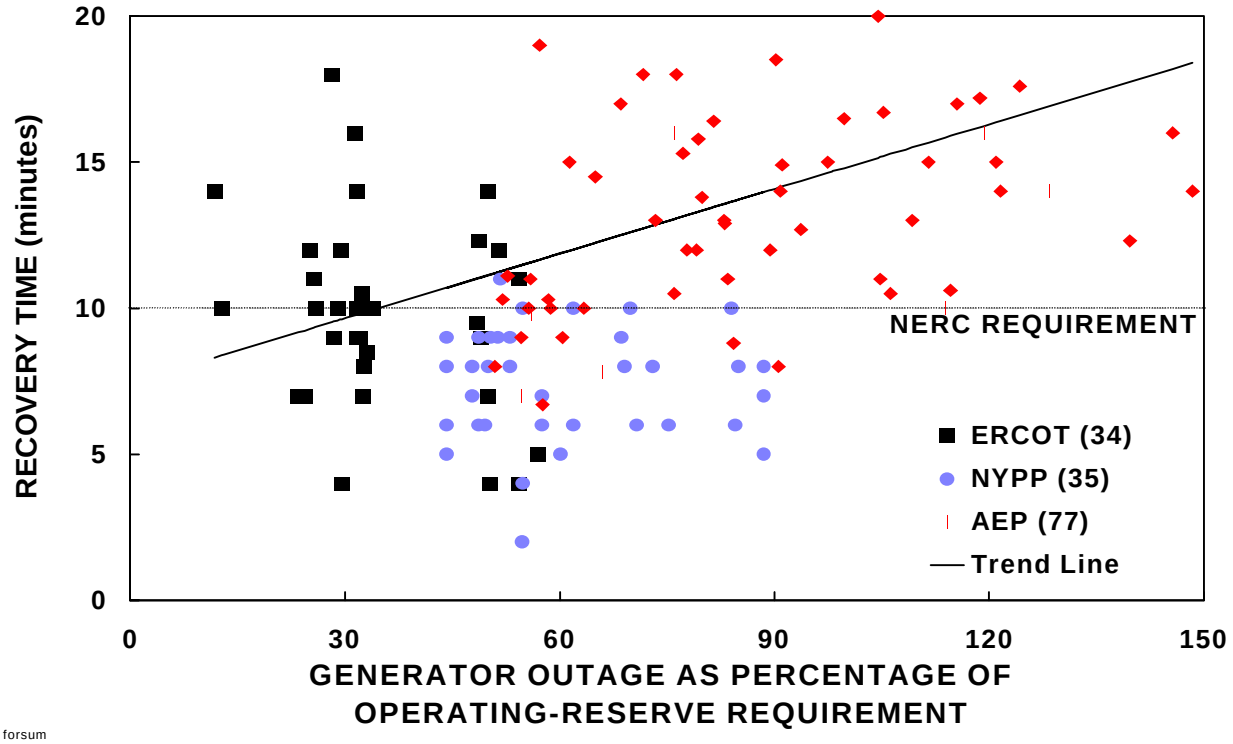

Figure 1Time to recover from a generator outage as a function of the outage magnitude relative to the minimum reserve requirement for three systems.

size affect recovery times. These other factors likely relate to the system conditions at the time of the outage, including the factors listed above. Finally, these data show that many outages exceed the NERC 10-minute requirement and will therefore lead to violations of DCS.

Bilke (1998) analyzed data from 208 outages in the Eastern and ERCOT interconnections. The average recovery time was about 10 minutes, with almost exactly half of the recovery times less than 10 minutes and half greater than 10 minutes. Compliance with the new DCS averaged about 80% in the Eastern Interconnection and about 90% in ERCOT. Bilke found a statistically significant relationship between outage size and recovery time; recovery time increases by 0.4 minutes for every 100-MW increase in outage size.

These results imply that the typical control area will have to increase the amount of reserves it carries by 20%. These required increases will impose financial penalties on suppliers and customers because generating capacity will be shifted from energy markets to operating reserves. Because of these costs, Bilke urges NERC to consider a DCS-compliance requirement of less than 100%. He also, suggests additional data collection to provide an "early warning system" on control-area performance and to better "police the interconnections." Alternatively, competitive markets for energy and ancillary services may motivate the industry to develop innovative and lower-cost ways to meet DCS, for example, through greater use of interruptible loads.

REGULATED COSTS AND PRICES

Today, operating reserves are generally sold under FERC-approved prices, which are based on embedded costs. The exceptions (discussed below) refer to the FERC-approved ISOs that plan to buy and sell reserves at market-based prices.

A typical process used to determine the price of operating reserves is as follows:

- Determine how much of the service is required. This is usually straightforward because it relies on reliability-council requirements. The minimum is usually based on the projected daily peak load or on the largest single contingency. Utilities rarely claim cost recovery for more reserves than the minimum required.[*] Historically, utilities usually carried more reserves than needed (especially during nonpeak hours) because of unit-commitment constraints.

- Identify the generating units that provide the services. The simplest approach is to assign a "slice-of-the-system" to the reserves, equivalent to assuming that all the generating resources in a utility's portfolio contribute proportionately to operating reserves. This approach ignores the fact that some units (in particular, nuclear units and low-fuel-cost coal units) almost never provide these services because they operate at maximum output. Thus, utilities may need to provide data showing which units actually provide each service. Unit ramp rate (MW/minute), the delay in response to a control-center request for a change in output, and other characteristics of generating units affect which ones are used to provide reserves. We have seen no filings that explicitly examine these characteristics.

- Identify the portion of capital costs associated with governors (spinning reserve only), boiler controls, and turbine controls. (These incremental costs would be shared between regulation and operating reserves.) Similarly, identify the capital costs associated with making a unit fast-start-capable for supplemental reserve.

- Identify the costs associated with unit commitment (spinning reserve only) and the incremental fuel plus operations and maintenance costs. Southern includes a unit-commitment cost of $47/MW-day in its derivation of the spinning-reserve charge. Encotech (1997) developed a computer model to estimate the heat-rate penalty associated with suboptimal turbine-valve operation for spinning reserve.

---

[*] A FERC (1997) administrative law judge noted that "the ECAR recommendations are 'minimums.' Thus, one cannot determine based solely on a reference to ECAR Document No. 2 what the aggregate should be for all of the three services [regulation, spinning reserve, and supplemental reserve] which Order No. 888 requires ... ."

Divide the total annual cost by an appropriate divisor (e.g., annual system peak or the average of the 12 monthly peaks). The result, in $/kW-year, is the annual cost of the reserve service.

Wisconsin Public Service Corp. (WPSC), in its open-access tariff, uses a conceptually attractive method for determining the annual capital costs of regulation, spinning reserve, and supplemental reserve. WPSC uses information from its energy-management system to determine, for each hour of a year, the status of each generating unit. It assigns units to the three ancillary services as follows:

Off and available: for combustion turbines only, assign to supplemental reserve with the number of MW so assigned based on the maximum output of the unit within 10 minutes. (This assignment could also apply to fast-start hydro units.)

On and under automatic control: assign to regulation based on the operating hours under automatic control. The capacity so assigned should probably be based on the smaller of (1) the maximum increase in output from the unit within 10 minutes (based on regulation ramp rate) and (2) the remaining headroom on unit (difference between maximum and current output levels).

On and under either manual or automatic control: assign to spinning reserve with the number of MW so assigned based on the smaller of (1) the maximum increase in output from the unit within 10 minutes (based on spin ramp rate) and (2) the remaining headroom on unit.

For spinning reserve, WPSC calculates for each unit the sum over the year of the number of MW available for spin to develop a MW-hour total for the year. The sum of these totals over all units yields the total MW-hour of spinning reserve for the system as a whole. Each unit's share of this total is used to determine the annualized capital cost. This approach could be refined to adjust the hourly values downward to reflect the total amount of spinning reserve required. The current approach credits each unit for all the capacity it makes available for spinning reserve regardless of whether or not the system needs that much capacity.

For supplemental reserve, WPSC calculates the MW capacity available from each of the combustion turbines for each of the hours that the unit was available but not online. Summing these MW-hour values yields a result analogous to that developed for spinning reserve. The method then proceeds as it does for spinning reserve. The WPSC approach ignores the possibility that supplemental reserve will be provided first from the unused headroom of units online and only later from the offline, fast-start combustion turbines.

These differences across generating units in contribution to different ancillary services is important because the embedded cost of each unit is different. For example, the annualized fixed costs for the WPSC generating units range from $16/kW-year (for a combustion turbine

that provides supplemental reserve) up to $216/kW-year (for a nuclear unit that provides no ancillary services).

COMPETITIVE-MARKET PRICING

Ultimately, competitive markets will likely develop for operating reserves. In competitive markets, historical costs are of little importance. Market-based prices will reflect the opportunity costs a unit incurs when it withholds capacity from energy markets to provide operating reserves, the heat-rate penalty a unit incurs by standing ready to respond rapidly, and the costs to redispatch other units to provide enough operating-reserve-capacity headroom. Additional costs incurred during reserve deployment will have to be covered as well. These costs include fuel use plus wear and tear caused by the required fast response. Finally, increased capital costs required to make the units responsive or fast start will be included in the reserve price. Costs are incurred both when the resource is standing by and when the service is provided. Prices, whether they explicitly differentiate between these costs or not, will have to recover both types of costs.

In California, the ISO conducts a day-ahead auction for several ancillary services, including spinning reserve, supplemental reserve, and replacement reserve. These auctions set the prices for each service for each of the 24 hours during the next day. These auctions are conducted after the California Power Exchange has closed its auction for the day-ahead energy markets. In addition, the ISO operates a real-time market for these services. The auctions are conducted sequentially, from regulation to spinning reserve to supplemental reserve and finally to replacement reserve. Capacity not acquired in one market can be rolled over into the next market.

California experienced several problems with these markets during the summer of 1998, primarily insufficient bidding and very high prices. As Wolak, Nordhaus, and Shapiro (1998) noted: "... the ISO's ancillary services markets do not yet operate in a manner consistent with workable competition. ... Ancillary service markets have exhibited extreme price volatility, even during periods when demand was unchanged for long periods of time. ... Prices for lower quality services such as replacement reserve routinely exceed the prices for higher quality services such as regulation. Often ancillary services capacity prices exceed both the power exchange and real-time energy price for the same hour." These problems occurred because (1) some generating firms were subject to cost-based price caps while others were allowed to earn market-based rates; (2) the amount of each service bought by the ISO was independent of the price; (3) the ISO was unable to purchase more of a valuable, but cheaper service and use it to displace some of a less valuable, but more expensive service; and (4) the ISO dispatch and settlement practices were not clear to many market participants.

Allowing generators to receive market-based rates in the energy markets while constraining them to cost-based rates in the ancillary-service markets creates perverse

incentives. Allowing some generators to receive market-based rates in the ancillary service markets only exacerbates this problem. The cost-based generators will remain in the energy market if the return there exceeds the cost-based return allowed in the ancillary-service markets. With the cost-based generators shunning the ancillary-service markets, supplies are tight and prices rise dramatically for the few generators with the authority to charge market-based prices.

The former tight power pools in the mid-Atlantic states (PJM Interconnection) and New England (ISO New England) are taking a different approach to acquisition of ancillary services. Although ISO New England plans to create markets for ancillary services, it also plans to retain much of its traditional unit-commitment and least-cost dispatch functions, managed through the ISO's computer programs. Thus, the New England plan is quite centralized in contrast to California's decentralized decision making and risk taking (in which the owners of generating units do their own unit commitment and dispatch and accept all the risks associated with those decisions).

The New England proposal, as of September 1998, called for the payment for operating reserves for "bid costs, lost opportunity costs, and production cost changes" (Cramton and Wilson 1998). It is not clear whether these "production cost changes" refer to the unit that is supplying the operating-reserve service or to other units that are redispatched to compensate for changes in output from units providing reserves. It is also unclear why the bids should be supplemented by ISO determinations of opportunity costs and production cost changes; perhaps the bidders should accept the risks of internalizing these factors into their bid offers. Also, the ISO New England formula for spinning-reserve payment includes a factor of 2, which implies that the generator owners are paid twice for the same service.

UNRESOLVED ISSUES

Technical Basis

NERC's new DCS is an important step in the direction of developing and implementing measurable, reliability-related performance standards. Unfortunately, only a limited analytical basis appears to exist for either the 10-minute-recovery-period requirement or for the minimum-operating-reserve requirements specified by the ten reliability councils. The DCS does not address the interconnection requirement for immediate response to arrest frequency decay. No basis appears to exist for the requirement that control areas calculate the DCS for only those outages that fall between 80 and 100% of the most severe single contingency. Limited analysis of data from three control areas shows that recovery times depend strongly on outage size as well as other factors (Fig. 1). Nor does the DCS appear to require the immediate response of spinning reserve; supplemental reserve may suffice to meet the DCS.

If the intent of the DCS is to maintain reliability at its current levels, then it may be appropriate to analyze existing performance in the three interconnections and to specify DCS requirements accordingly. (We have seen nothing to suggest that current outage-response performance is unacceptable.)

Such rules of thumb may have been sufficient in the past. In the future, customers will be required to pay explicitly for these operating reserves (because their costs will no longer be buried within the bundled electric rate they formerly paid). Customers and power marketers will want clear evidence that they are getting the service for which they are paying and that this service is required to maintain bulk-power reliability. Therefore, they will want a say in how these standards are set. And generators that impose less of an operating-reserve burden on the system will want to be compensated accordingly; alternatively, these units will require less reserve capacity.

Operating-reserve requirements could be based on either probabilistic or deterministic calculations. Ideally, the reliability councils would conduct both types of analyses, testing the results against different sensitivities related to unit-specific forced-outage rates, capacity costs, transmission constraints, and other factors. In general, these requirements are currently based primarily on the magnitude of the largest single contingency. The thinking behind this approach is that the system must be able to withstand such a contingency regardless of the probability of its occurrence.

An alternative approach, which merits additional consideration, would focus on the statistics of outages, their severity, and their consequences. In such a probabilistic approach, the performance of individual generators would figure prominently in the determination of the minimum amount, type (capital vs operating costs), and location of operating reserves required.

Historically, it may not have been important to recognize the frequency of forced outages at individual generators because utilities provided operating reserves from their portfolio of generating resources and sold this service as part of the bundled electricity product. In the future, with generation unbundled from transmission and with generation increasingly competitive, system operators may need to recognize differences in reliability among generators. These differences could affect the amount of reserves required, how reserves are deployed, and the allocation of costs among generators. Figure 3 shows the number of forced outages for coal-fired units between 600 and 800 MW for 1996 (Curley 1997). Although, on average, these 120 units experienced 9 outages a year, two units had 0, while at the other extreme, five units had 20 or more outages.

The outage frequency affects the cost of providing operating reserves and should influence the types of generators chosen to provide reserves. For example, a generator that trips often requires the operating reserves to deploy frequently. This generator might be better

served by reserves with low deployment cost even if they have high standby costs. An extremely reliable generator that trips only rarely, on the other hand, might be better served by a resource with low standby costs but high deployment costs, such as automatically interruptible load or fast-start generation.

Finally, NERC and the reliability councils should consider the feasibility of eliminating the minimum operating-reserve requirements. Allowing each system operator to determine how much and what type of reserves to carry—so long as the control area meets its DCS requirements—might yield lower-cost ways to maintain reliability.

Paying for Reserves

The current systems for assigning operating reserves do not distinguish among the performance of *individual* generators. In the future, the system may change in two ways.

- The amount and type of reserves required at any given time should depend on which generators are online. If that mix of generators is highly reliable, the operating-reserve requirements will be modest. On the other hand, if some of those generators have frequent forced outages, the overall operating-reserve requirement may be higher and may consist of a different mix of generators (e.g., units with high capital costs but low operating costs).

- Although customers will ultimately pay for operating reserves, in the first instance suppliers may pay. Specifically, each supplier could be assigned an amount of operating reserves to provide or obtain as a function of its individual unit performance. And when a particular unit trips offline, that unit could be responsible for any extra payments to generators that provided energy during the 30- or 60-minute period that such reserves operate.

In a competitive market, generating units that are highly reliable (e.g., those on the left side of Fig. 2) might be required to provide or pay for different or less operating reserves than would the units that experience frequent outages. This economic signal would provide the appropriate incentive to generation owners, encouraging them to undertake the amount of maintenance that would just balance the higher cost of providing more and more-expensive operating reserves. In addition, when an outage occurs and operating reserves are called upon, the generator responsible for the outage would pay the incremental costs of the units that responded to the outage (i.e., the additional fuel plus operating costs beyond those associated with the spot-market price for that hour). This pricing approach would eliminate subsidies among generators and would provide further incentives to generator owners to maintain high availability levels at their units. Finally, competitive markets might discipline generators with poor reliability records by requiring them to carry larger amounts of backup supply than those units that are highly reliable.

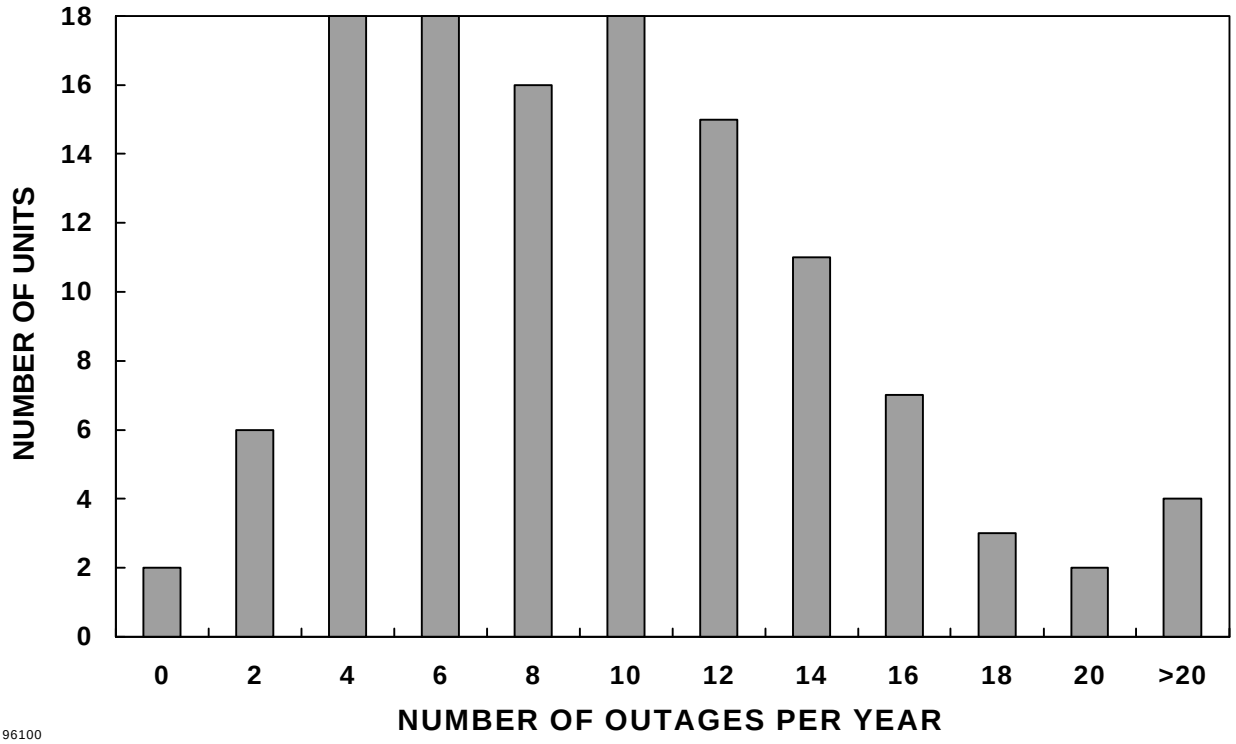

Figure 2Number of forced outages in 1996 for 120 coal-fired generating units between 600 and 800 MW in size.

Future Data Needs

We found it difficult to obtain data on forced outages. Utilities, power pools, and regional reliability councils are often unable to readily provide information on the frequency and consequences of outages. Control centers often cannot provide data on the number of units online at the time of the outage, the amount of operating reserves online, and whether the unit that tripped offline was also providing operating reserves. Because a competitive electricity market will likely insist on technically defensible reserve requirements, NERC should collect more data from individual utilities and ISOs along the following lines: unit output (MW) immediately before the outage, unit capacity (MW), system load at the time of the outage (MW), system operating reserves at the time of the outage (MW), number of units online at the time of the outage, and the time to return ACE and frequency to their precontingency values.

Mixing Functions

Currently, operating reserves are often used to protect against load-forecast errors as well as supply outages. It is inappropriate to assign operating reserves to two disparate

functions even when the services are provided by the same generating units.* The load-forecast errors and generator forced outages represent very different phenomena, with the former being a commercial function and the latter a bulk-power-reliability function. The cost of maintaining reserves to protect against generator outages should, as discussed above, fall on those generators responsible for the operating-reserve requirement (e.g., those with frequent forced outages or large unit sizes). The cost of maintaining reserves to protect against load-forecast errors should be assigned to those customers or scheduling coordinators that forecast poorly; these costs should appear in the load-following, energy-imbalance, or backup supply services, not in operating reserves. Finally, the times for deploying reserves for these two functions differ substantially. Reliability reserves must respond rapidly and achieve their full output within 10 minutes, usually with no advance warning. Responding to load-forecast errors is a much more gradual process that occurs over several hours.

Mixing the responses to forecast errors and contingencies is troubling for two additional reasons. First, when operating reserves are deployed to correct for load-forecast errors, the system is exposed to increased risks associated with forced outages. All users of the bulk-power system share this risk without their approval or knowledge, regardless of which users were responsible for the load-forecast errors that led to the increased risk. Second, the mixing of functions within one service invites gaming. Individual market participants will lean on the system as much as possible, to shift their costs to other participants.

Although FERC defined separate operating-reserve and regulation services, NERC and regional reliability requirements often include regulation (the use of generating units to follow minute-to-minute fluctuations in system load). These services should be treated separately. Generators that provide regulation grant the system operator the right to vary the outputs of their units up or down, from minute to minute, within a specified range ($\pm x$ MW with a ramp rate of y MW/minute). Generators that provide operating reserves grant the system operator the right to increase (but not decrease) the output from their units when a major disturbance occurs, often at a faster ramp rate than that associated with regulation. A unit providing operating reserves can expect to be called upon to provide those reserves a few times a month. Thus, the operation of generators for regulation differs substantially from the operation of those units for operating reserves. Combining these two disparate functions is inappropriate.

CONCLUSIONS

*Although these services (operating reserves, load following, and regulation) may be provided by the same generating units, their requirements, costs, prices, and billing to customers should be considered separately. NERC's definition of contingency reserves as the portion of operating reserves assigned to reliability functions is a step in this direction.

Creating competitive markets for ancillary services is essential because of the close link between these services and the basic energy service. As Bohn, Klevorick, and Stalon (1998) wrote, "... it is fundamental to recognize that the capacity available for the ISO [ancillary service] markets and for the PX [energy] markets comes from the same generating capacity. Capacity sold in one market means less capacity that can be sold in other markets, thereby driving up prices in the latter. Therefore, we would expect a close relationship among the different markets."

Although FERC views operating reserves as a reliability service only, NERC and most utilities combine reliability and commercial functions in the same service. This combination is unwise because protection against major generation and transmission outages differs both in purpose and in operation (speed of response and advance notice) from adjustment of generation for load-forecast errors.

NERC's performance requirements to protect the grid from problems associated with major generator or transmission outages need either to be revised or more fully documented. In particular, NERC's Disturbance Control Standard lacks any published justification and seems to bear little relationship to the reserve services that are intended to meet the standard.

Finally, because operating reserves are essential for bulk-power reliability, the electricity industry must develop market structures and rules that encourage efficient supply, acquisition, and use of this service.

Table 1 summarizes these and other unresolved issues related to operating reserves in increasingly competitive bulk-power markets.

Table 1. Key unresolved issues for operating reserves

What functions should be included in this service? In particular, should the generating capacity needed to adjust for load-forecast errors be shifted from operating reserves to the load-following service?
What is the technical basis for NERC's Disturbance Control Standard?
How does the Disturbance Control Standard support the distinctions between spinning and supplemental reserve? That is, how does the immediate response of spinning reserve improve compliance with the Disturbance Control Standard?
What is the technical basis for today's minimum-operating-reserve requirements specified by the regional reliability council? What is the basis for the differences among regions in these requirements? What data and analysis are needed to improve the technical justification for these requirements?
What data should NERC and the regional reliability councils collect, and what analyses should they do to support the requirements for operating reserves?

REFERENCES

T. Bilke 1998, *Maintaining Historic Reliability Under the NERC Disturbance Control Standard*, draft, Wisconsin Electric Company, Milwaukee, WI, February.

R. E. Bohn, A. K. Klevorick, and C. Stalon 1998, *Report on Market Issues in the California Power Exchange Energy Markets*, Market Monitoring Committee of the California Power Exchange, Alhambra, CA, August 17.

P. Cramton and R. Wilson 1998, *A Review of ISO New England's Proposed Market Rules*, prepared for ISO New England, Holyoke, MA, September 9.

M. Curley 1997, personal communication, North American Electric Reliability Council, Princeton, NJ, September.

Encotech Engineering 1997, *Cost of Providing Ancillary Services from Power Plants, Operating Reserve—Spinning*, TR-107270-V4, Electric Power Research Institute, Palo Alto, CA, July.

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

E. Hirst and B. Kirby 1998, *The Functions, Metrics, Costs, and Prices for Three Ancillary Services*, Edison Electric Institute, Washington, DC, October.

Interconnected Operations Services Working Group 1997, *Defining Interconnected Operations Services Under Open Access*, EPRI TR-108097, Electric Power Research Institute, Palo Alto, CA, May.

N. Jaleeli and L. S. VanSlyck 1997, *Control Performance Standards and Procedures for Interconnected Operation*, EPRI TR-107813, Electric Power Research Institute, Palo Alto, CA, April.

North American Electric Reliability Council 1996, "Control Performance Standard and Disturbance Control Standard Frequently Asked Questions," Princeton, NJ, November.

North American Electric Reliability Council 1997, *NERC Operating Manual*, Princeton, NJ, December.

North American Electric Reliability Council 1998, *Policy 10 – Interconnected Operations Services*, draft, Princeton, NJ, April.

U.S. Federal Energy Regulatory Commission 1997, *Initial Decision: Northern Indiana Public Service Company, Inc.*, Docket No. ER96-399-000, Washington, DC, May 9.

R. Vice 1998, personal communication, System Operations Department, Southern Company Services, Birmingham, AL, February.

Western Systems Coordinating Council 1997, *WSCC Operating Reserve White Paper*, Salt Lake City, UT, June 4.

F. Wolak, R. Nordhaus, and C. Shapiro 1998, *Preliminary Report on the Operation of the Ancillary Services Markets of the California Independent System Operator*, Market Surveillance Committee, California Independent System Operator, Folsom, CA, August 19.

D:\TEXT\REPORTS\Technical & Market Issues for Operating Reserves.wpd