BULK-POWER BASICS: RELIABILITY AND COMMERCE

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

January 24, 2000

1. INTRODUCTION

The restructuring of the U.S. electricity industry offers important opportunities and challenges to the environmental community. The opportunities arise because the traditional, vertically integrated utilities are being broken up into their competitive and regulated components; environmentally benign demand and supply resources will have greater opportunities to compete for market share in this new world. The challenges arise because bulk-power reliability and commerce are tightly integrated; environmental advocates must understand how bulk-power systems are planned and operated, under both normal and contingency conditions, to participate effectively in commercial markets.

This paper explains the basics of bulk-power systems and how the institutions that affect reliability and commerce are changing. These institutional changes are a consequence of the restructuring now underway throughout the United States. Although the transition from yesterday's vertically integrated utilities to tomorrow's industry dominated by a diversity of competitive and regulated entities is far from complete, we offer our model of the future industry structure (Fig. 1). This possible end-state structure includes three regulated entities (shown as rectangles) and three competitive entities (shown as ovals).

Generation encompasses a variety of independent companies that own one or more generating stations. These competitive entities sell power and ancillary services through a variety of arrangements, such as long-term bilateral contracts, day-ahead energy markets run by power exchanges, or on a real-time (intrahour) basis through spot markets run by the system operator. They earn or lose money on the basis of their ability to operate and maintain their plants inexpensively and their ability to market their electrical output skillfully.

System operators are either nonprofit or for-profit entities that may or may not be combined with transmission companies; that is, what the U.S. Federal Energy Regulatory Commission (FERC) calls regional transmission organizations (RTOs) may own and operate transmission networks or may only direct the operation of transmission networks. The RTOs perform services analogous to those performed by the air-traffic control centers at each airport. The RTOs, regulated by FERC, are monopolies, with one and only one RTO in each electrical region. The RTOs are responsible for short-term reliability [what the North American Electric Reliability Council (NERC) calls security]. In addition, the RTOs ensure that all qualified entities (e.g., generators, power marketers and brokers, and customers) have comparable, nondiscriminatory, and open access to the transmission grid.

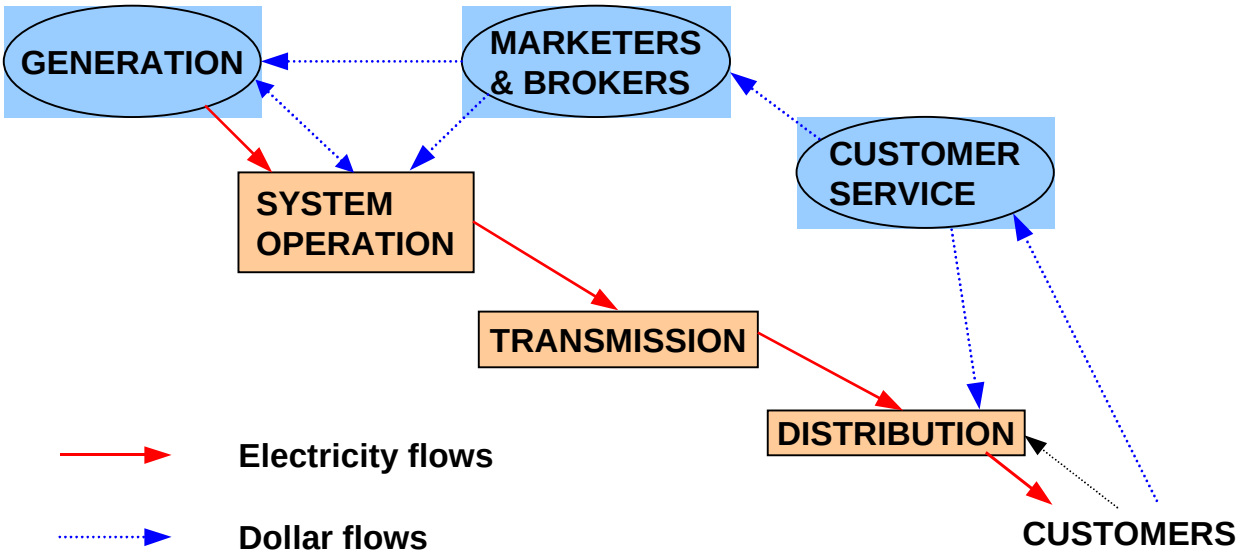


Fig. 1. Possible structure of the U.S. electricity industry. Rectangles are entities that are regulated, while ovals are unregulated entities.

Transmission companies, also regulated by FERC, own and operate transmission systems. These entities, under the guidance of the RTO in that region, maintain and operate the system and may also be responsible for constructing new transmission equipment (e.g., transformers, substations, and high-voltage lines).

Distribution entities, regulated by state public utility commissions (PUCs), are the low-voltage analogs to the transmission entities. These wires companies move power from the transmission system to retail customers. Although most generators are connected to the electrical system at the high-voltage (transmission) level, distributed resources are more likely to be connected to the distribution system. Because this paper focuses on bulk-power issues, distribution issues are not discussed here.

Power marketers and brokers are entities that buy power from generators and sell power and related services to wholesale and retail (end use) customers. Customer-service entities offer other services to electricity consumers, such as energy efficiency, bill consolidation, and management of a customer's total energy supply (e.g., natural gas plus electricity). The distinction between power marketers and brokers, on one hand, and customer service companies, on the other, will likely change and diminish over time. Like generation, these companies are largely unregulated.

Customers, within this future industry structure, will buy electricity and related services from power marketers, brokers, and energy-service companies. Their electric bills may be unbundled, with separate charges for generation, transmission, ancillary services, distribution, and customer service. Only the transmission and distribution parts of their bills will continue

to be regulated by government agencies; their energy, ancillary-services, and customer-services charges will be largely unregulated.

The complexity associated with possible bulk-power structures and operations derives less from the basic market and reliability functions and more from the multiplicity of combinations that can occur. Markets can be distributed (i.e., bilateral contracts) or centralized; long- or short-term; covering energy, transmission, ancillary services, or combinations of all three products; and markets and system operation can be managed together or separately by investor-owned or nonprofit entities.

Reliability and markets are tightly coupled. Bulk-power reliability cannot be easily and unambiguously defined (see Exhibit 1). But we know when the lights are off. A reliable electric system is one that allows for few interruptions of service to customers. Outages can be defined in terms of their number, frequency, duration, and amount of load (or number of customers) affected. Equally important, but much more difficult to quantify, is the value of loss of load.*

Although generation and transmission failures cause only a small fraction of the power outages, their economic and societal consequences can be much higher than those associated

Exhibit 1. NERC's Reliability Definition

The North American Electric Reliability Council (NERC), the primary guardian of bulk-power reliability, was established in 1968. NERC's creation was a direct consequence of the 1965 blackout that left almost 30 million people in the northeastern United States and Ontario, Canada without electricity.

NERC defines reliability as "the degree to which the performance of the elements of [the electrical] system results in power being delivered to consumers within accepted standards and in the amount desired." NERC's definition of reliability encompasses two concepts, *adequacy* and *security*. Adequacy is defined as "the ability of the system to supply the aggregate electric power and energy requirements of the consumers at all times." It defines security as "the ability of the system to withstand sudden disturbances."

In plain language, adequacy implies that there are sufficient generation and transmission resources available to meet projected needs plus reserves for contingencies. Security implies that the system will remain intact even after outages or other equipment failures occur.

*A 10-minute power outage in a residence is an annoyance because someone has to reset the digital clocks but imposes only small economic costs. But a similar outage for a computer-chip manufacturer might entail the loss of millions of dollars of output.

with distribution outages. Bulk-power outages generally affect many more customers and are much more difficult to recover from than is true for distribution outages.

The rest of this paper proceeds as follows. Section 2 explains the unique features of electric systems and how these features affect operation of bulk-power systems. Section 3 describes the key bulk-power institutions and how they are changing as the electricity industry is restructuring. Section 4 discusses real-time pricing, a key element of competitive electricity markets. Section 5 discusses transmission congestion and locational prices, a topic of increasing importance as the number of bulk-power transactions increases. Section 6 discusses planning and investment in new generation and transmission facilities. And Section 7 summarizes the key points developed here.

2. UNIQUE FEATURES OF ELECTRICITY

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

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

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

These two unique characteristics lead to four reliability consequences with practical implications that dominate power system design and operations:

- Every action can affect all other activities on the grid. Therefore, the operations of all bulk-power participants must be coordinated.

- Cascading problems that increase in severity are a real problem. Failure of a single element can, if not managed properly, cause the subsequent rapid failure of many additional elements, disrupting the entire transmission system.

- The need to be ready for the next contingency, more than current conditions, dominates the design and operation of bulk-power systems. It is usually not the present flow through a line or transformer that limits allowable power transfers, but rather the flow that would occur if another element fails.

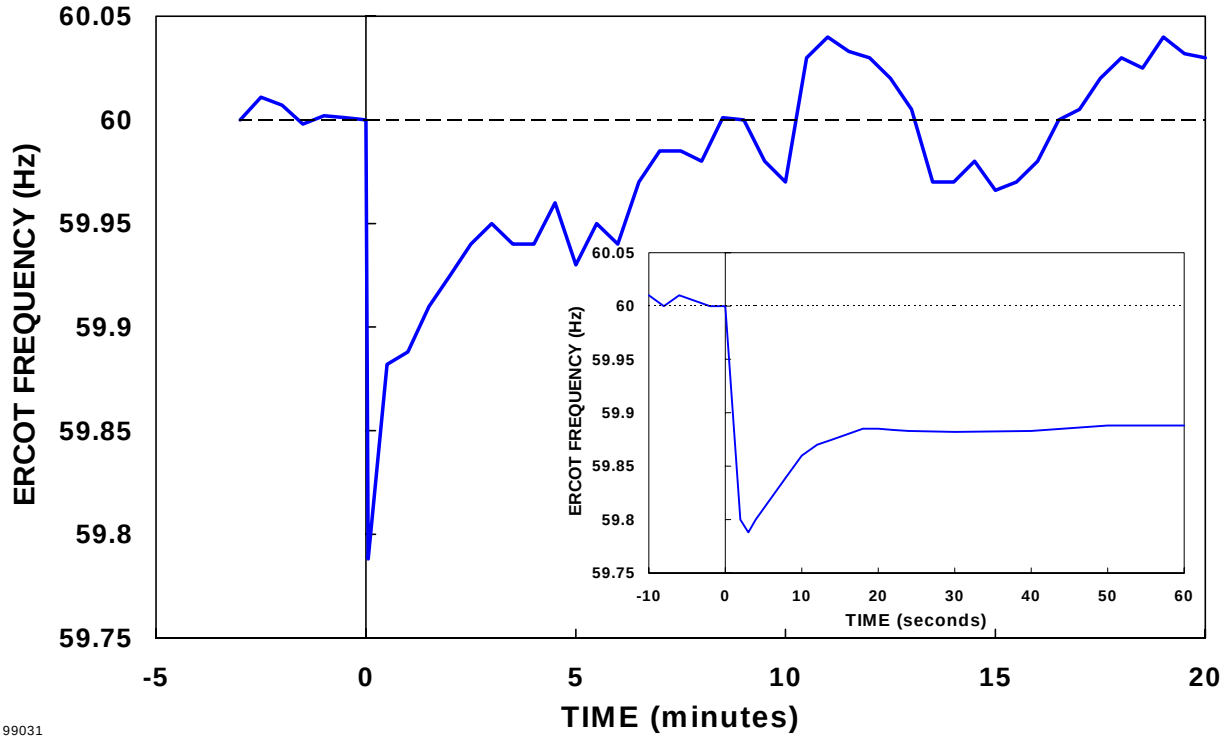


Fig. 2. **Interconnection frequency before and after the loss of a 653-MW generator. The inset shows frequency for the first minute after the outage, and the larger figure shows frequency for the first 20 minutes after the outage.**

- Because electricity flows at the speed of light, maintaining reliability often requires that actions be taken instantaneously (within fractions of a second), which requires computing, communication, and control actions that are automatic.

Lightning provides an example of situations in which automatic responses are required. When lightning strikes a transmission line, breakers at both ends of the line sense the high current and open automatically. Within a fraction of a second (enough time for the fault to clear), the breakers close again and power, once again, flows through the line. Because this process occurs so quickly, it takes place with no human intervention.

Responding to a major generation outage provides another example of how the electricity industry responds to these unique features. Figure 2 illustrates how the electric system operates when a major generating unit suddenly fails. Prior to the outage, system frequency is very close to its 60-Hz reference value. Generally, within a second after the outage occurs, frequency drops, in this case to 59.79 Hz. The frequency decline is arrested primarily because many electrical loads (such as motors) are frequency responsive; that is, their demand varies with system frequency. Once the frequency decline exceeds the deadband of the generator governors, the governors at those generators so equipped sense the frequency decline and open valves on the turbines, which rapidly increases generator output. This governor

response accounts for the initial increase in frequency during the first several seconds after the outage occurs, as shown in the Fig. 2 inset. At this point, the generating units that provide contingency reserves, in response to signals from the control center, begin to increase output. More fuel is added to the boiler, leading to a higher rate of steam production, which leads to higher power output. In this example, the system worked as it was intended to, and frequency was restored to its pre-contingency 60-Hz reference value within the required 10 minutes (at 8.5 minutes).*

3. KEY BULK-POWER INSTITUTIONS

Several different types of organizations oversee, operate, and participate in bulk-power markets and reliability. These entities, many of which did not exist a few years ago, range from private companies to government agencies, include control-area operators (often called system operators), Interconnections, security coordinators, regional reliability councils, NERC, and FERC.

The North American electric system is divided into three Interconnections (Fig. 3): Eastern, Western, and the Electric Reliability Council of Texas (ERCOT, which covers most of Texas). Within each Interconnection, all the generators operate at the same frequency as essentially one machine connected to each other and to loads primarily by AC lines. The Interconnections are connected to each other by a few DC links. Because these DC connections are limited, the flows of electricity and markets are much greater within each Interconnection than between Interconnections.

The fundamental entity responsible for maintaining bulk-power reliability is the control area. NERC defines control areas as: "An electric system or systems, bounded by interconnection metering and telemetry, capable of controlling generation to maintain its interchange schedule with other Control Areas and contributing to frequency regulation of the Interconnection." Control areas are linked to one another to form Interconnections. Each control area seeks to minimize any adverse effect it might have on other control areas within the Interconnection by (1) matching its schedules with other control areas (i.e., how well it matches its generation plus net incoming scheduled flows to its loads) and (2) helping the Interconnection to maintain frequency at its scheduled value (nominally 60 Hz).

Today's approximately 150 control areas are operated primarily by utilities, although a few are run by independent system operators (ISOs). Control areas vary enormously in size, with several managing less than 100 MW of generation and, at the other end of the spectrum, PJM (Pennsylvania-New Jersey-Maryland Interconnection) and California each managing about 50,000 MW of generation. Control areas are grouped into regional reliability councils,

*The NERC Security Committee recommended, in November 1999, that the allowable disturbance-recovery period be extended from 10 to 15 minutes.

of which there are 10 in the 48 contiguous states, most of Canada, and a small portion of Mexico (Fig. 3). These reliability regions, in turn, are parts of the three Interconnections.

As bulk-power markets become more competitive, the institutions that oversee and manage reliability and commerce are changing. Historically, utilities, and only utilities or their power-pool aggregations, owned generation and transmission and therefore operated control areas. Within the last few years, several ISOs that own neither generation nor transmission, have taken over these functions in California, New England, New York, and the mid-Atlantic (PJM) region. Several other ISOs and Transcos* are now under development. In December 1999, FERC issued a major rule (Order 2000) on RTOs. The order requires utilities to file reports with FERC by October 2000 with a proposal for joining an RTO or an explanation of why the utility cannot join such a regional organization. The Appendix summarizes Order 2000.

Recently, a large power marketer proposed to create three generation-only control areas inside the boundary of an existing utility control area, presumably to gain competitive advantages that it sees in today's control-area-operator functions and privileges. These advantages include the ability to purchase power day ahead and then wait until the operating hour to decide where to send (or sell) that power, the ability to use the payment-in-kind feature

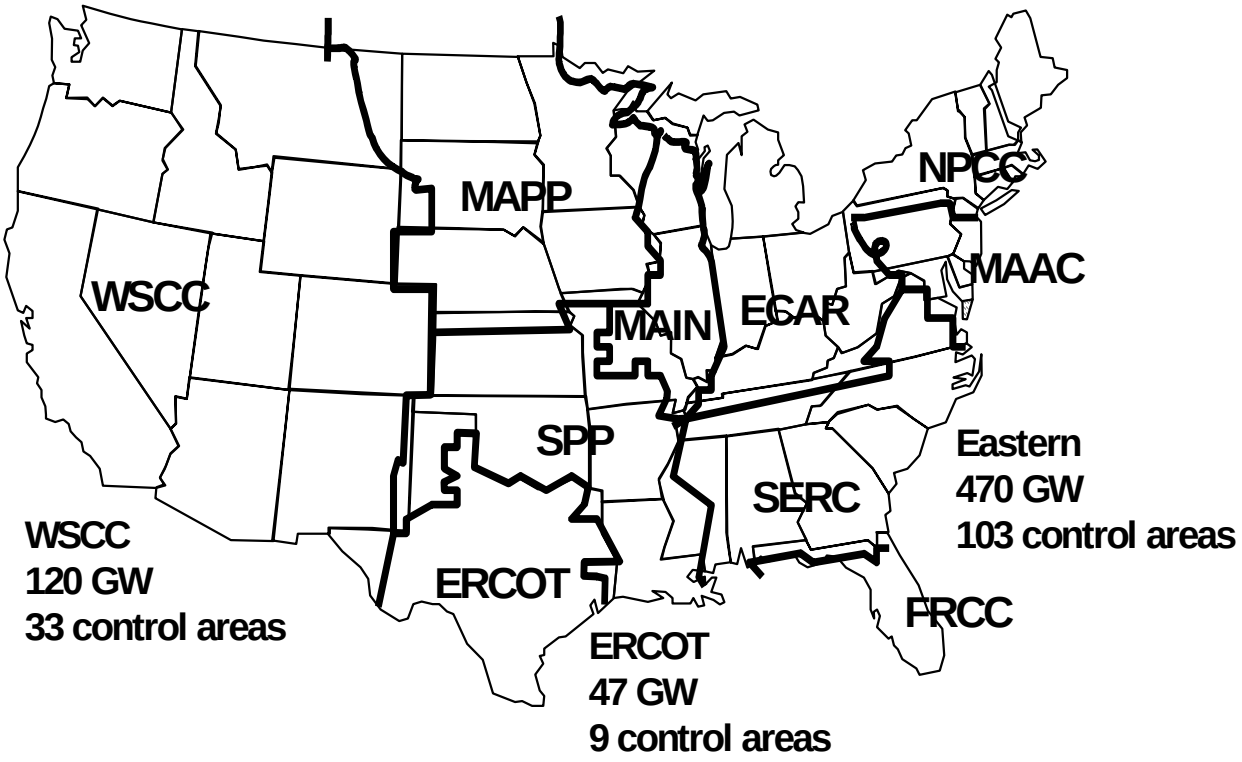


Fig. 3. U.S. map showing the locations of the three Interconnections and 10 regional reliability councils.

*Transcos differ from ISOs in that they own, as well as operate, transmission facilities. In addition, all the existing Transco proposals will be for-profit entities, whereas all the current ISOs are not-for-profit entities.

of inadvertent interchange rather than the cash payment required for energy imbalance (Exhibit 2),[*] the ability to avoid having transactions cut by transmission loading relief requirements, and access to reliability information that may yield commercial advantages.

Not surprisingly, the utilities that now operate control areas are opposed to the creation of such new enterprises. In response to these requests to create new control areas, NERC set up a Control Area Criteria Task Force to write reliability policies and standards that apply to the "root level entity." This effort will define reliability entities (which will likely extend beyond control areas to include, for example, security coordinators and regional reliability councils) and ensure that no reliability entity has competitive market advantages.

Table 1 shows the range of functions that a control-area operator, ISO, or Transco might provide. The time frames over which these actions occur range from fractions of a second (e.g., operation of automatic protection devices to clear faults, protect people, or protect expensive equipment from damage) to years (planning and building new transmission facilities). Unambiguously labeling these activities *reliability* or *commercial* is very difficult. Most functions affect both reliability and commerce. Some functions that were traditionally performed by the vertically integrated utilities, such as day-ahead unit commitment and real time economic dispatch of generating units, may not be done by control-area operators in the future because these functions are primarily commercial and not reliability related.

The "higher level" bulk-power entities are changing as well. FERC is the federal agency with jurisdiction over bulk-power markets, including interstate transmission systems. In response to the Energy Policy Act of 1992, FERC implements policies (including its Orders 888 and 889) to assure that the owners and operators of transmission facilities under the agency's jurisdiction provide nondiscriminatory service to all power suppliers in wholesale power markets.[#] Historically, FERC has not had to involve itself with regulating reliability functions. Increasingly, some parties are calling on FERC to exercise its authorities by addressing reliability issues that intersect with the commercial needs of the industry.

Electric utilities established NERC in 1968 as a voluntary membership organization as an alternative to government regulation of reliability. NERC develops standards, guidelines, and criteria for assuring system security and evaluating system adequacy. NERC is funded by the 10 regional reliability councils, which adapt NERC rules to meet the needs of their regions. NERC and the regional councils have largely succeeded in maintaining a high degree of transmission-grid reliability throughout North America. Historically, the reliability councils

---

[*]In July 1999, a utility in the ECAR region "borrowed" large amounts of energy from the Eastern Interconnection during high-priced periods. Rather than pay the hourly spot price for this energy, the utility was able to repay the energy in-kind over the next 30 days, in accordance with NERC policy on inadvertent interchange.

[#]FERC's jurisdiction includes investor-owned utilities only. The roughly 35% of U.S. transmission plant owned by municipal, rural cooperative, federal, and Texas (within ERCOT) utilities is not subject to FERC authority. However, many of these entities voluntarily comply with FERC orders.

Exhibit 2. Inadvertent Interchange and Energy Imbalance

A control area cannot, and indeed need not, balance generation to load perfectly. NERC Control Performance Standards (CPS) 1 and 2 set limits on the allowable error between a control area's load and generation. The hourly imbalance between generation and load is called inadvertent interchange, which NERC defines as: "The difference between a Control Area's net actual interchange and net scheduled interchange." Control-area operators do not pay cash for imbalances. Instead they repay imbalances in kind. The only restrictions on repayment are that they must be repaid (1) within 30 days and (2) during the same time period (either onpeak or offpeak, with onpeak defined as a 16-hour period during each nonholiday weekday and offpeak defined as the remaining hours during the week).

The transmission-customer analogue to inadvertent interchange is energy imbalance, which NERC defines as the "energy correction for any hourly mismatch between a transmission customer's energy supply and the demand served." According to FERC's Order 888, transmission customers with an hourly imbalance that falls outside a ±1.5% or 2-MW deadband must pay for their imbalances.

The discrepancy in treatment between inadvertent interchange and energy imbalance must be resolved in the long term. Ultimately, in our view, all imbalances, whether created by system operators or individual market participants, will be settled hourly (or intrahourly) on the basis of the then-current market price of electricity (see Section 4).

have functioned without external enforcement powers, depending on voluntary compliance with standards. NERC is now in the process of converting its system from one in which peer pressure encouraged compliance with voluntary standards into one in which compliance is mandatory and violations are subject to penalties (including fines). Absent federal legislation requiring compliance with reliability standards, NERC has limited ability to enforce its reliability rules.

NERC is also expanding greatly the representation from all industry sectors on its committees and the Board of Trustees; in particular, NERC recently expanded its Board to include nine new members who are not affiliated with any market sector. Upon passage of federal reliability legislation authorizing creation of a national reliability organization, NERC's industry-affiliated board members will resign, leaving a board that is entirely free of commercial interests in electricity markets. At that time, the name of the organization will be changed to NAERO, the North American Electric Reliability Organization.

Table 1. Bulk-power commercial and reliability functions

Long term (1 to 10 years)

Produce load and resource forecasts
Prepare plans for additional generation
Prepare plans for additional transmission
Ensure that sufficient generation is constructed to meet forecast peak demands and required installed-capacity margins
Ensure that sufficient transmission is constructed for reliability and to promote commerce (i.e., reduce congestion cost effectively)
Set interconnection requirements for generation and load
Set metering and communications requirements and standards for generation and load

Intermediate term (1 to 24 months)

Coordinate transmission maintenance schedules
Coordinate generation maintenance schedules
Develop and test system restoration plans for recovery from major outages

Scheduling (1 to 7 days)

Provide information on system conditions and day-ahead forecasts to market participants
Determine whether proposed schedules can be met without violating security constraints
Prepare final schedules to manage (avoid) congestion
Operate FERC approved Open Access Same Time Information System (OASIS) for transmission reservations and pricing
Implement FERC-approved transmission tariff
Establish transmission operating limits
Develop least-cost commitment schedules for generating units
Manage accounting, billing, and settlements

Operations (real time to hours)

Oversee and monitor physical operation of transmission, i.e., collect and analyze data on generator output, loads, transmission flows, voltages, and frequency
Order generation and transmission changes to maintain voltages, frequency, transmission loadings, and generation/load balance within required ranges
Operate online generating units to minimize operating costs
Act to prevent outages and to protect equipment (including congestion management)
Respond to disturbances (e.g., lightning or the sudden loss of generation)
Acquire and dispatch resources (including load) for ancillary services
Coordinate outages and returns to service

Until a few years ago, FERC and NERC operated on parallel tracks with little interaction needed between the two institutions. FERC oversaw bulk-power commerce, NERC oversaw bulk-power reliability, and there was little interaction between commerce and reliability. Unbundling generation from transmission and creating competitive markets for electricity are dramatically changing this situation. The industry now recognizes that reliability and commerce are tightly integrated. Increasingly, FERC receives cases in which market participants complain that NERC reliability rules, their implementation, or both competitively disadvantage them. NERC recently established a Market Interface Committee as a complement to its long-standing Security (Operations) and Adequacy (Engineering) Committees.

State public utility commissions, with few exceptions, will likely play little role in bulk-power operations, reliability, or commerce. FERC clearly has jurisdiction over bulk-power commerce, and the physics of electricity are largely regional and not local. While states may seek to preserve a role for themselves in the establishment and operation of RTOs, ultimately they will serve more as advisors than as regulators and managers.

In response to recent NERC requirements, 22 Regional Security Coordinators coordinate within the reliability regions and across the regional boundaries. These security coordinators conduct day-ahead security analysis, analyze current-day operating conditions, and implement NERC's Transmission Loading Relief (TLR) procedures to mitigate transmission overloads. Many of today's system-operation and security-coordination functions are managed by investor-owned utilities; others are run by federal power agencies, ISOs, or regional reliability councils. (The distribution of these security coordinators demonstrates the influence of the traditional vertically integrated investor-owned utilities, which account for 13 of the 22 Security Coordinators.)

4. REAL-TIME PRICING

HOURLY PRICES

Electricity prices are volatile, perhaps more volatile than for any other commodity, and will become more so as competition increases. They are so variable because:

- Generators differ substantially in their costs to produce electricity (e.g., the running costs for hydro and nuclear units are typically well below $10/MWh, while the cost for an old combustion turbine might be $100/MWh or more);

- System loads vary substantially from hour to hour (e.g., by a factor of two to three during a single day);

- Electricity cannot easily be stored and therefore must be produced and consumed at the same time;

- Intertemporal constraints are such that at certain low-load hours the price can be zero or negative because it costs more to turn a unit off and turn it on again later than to keep it running; and

- When unconstrained demand exceeds supply the price is set by consumer demand at a level above the running cost of the most expensive unit then on line.

One of the key effects of restructuring will be widespread availability of real-time (hourly) prices to retail customers. This statement does not mean that all retail customers will be required to buy electricity in this fashion; it does mean that customers will have the *option* to do so. Providing some customers with information on the real-time costs of their electricity consumption will have profound effects on the electricity industry and its operations. In particular, this information will encourage some customers to change their electricity use in real time in ways that will improve bulk-power reliability and lower electricity costs for all consumers.

The fraction of customers that will choose to face real-time prices will likely grow over time as the larger customers gain experience (and save money) with this approach (see Exhibit 3). In addition, the costs, diversity, scope, and quality of the metering, communications, and control technologies required to enable this approach are all improving with time. Thus, an economical decision today to not install the equipment necessary to participate in real-time markets might be reversed in a few years because the costs have dropped sufficiently.

Spot prices are likely to be highly volatile. For the one-week period shown in Fig. 4, the ratio of the maximum-to-minimum price of electricity was five, about triple the ratio for hourly loads. During the summer of 1997, the ratio of maximum-to-minimum electricity prices in the mid-Atlantic region was much higher, about 20. These large hour-to-hour differences in electricity price provide substantial opportunities to make money by selling electricity at the right times (when prices are high rather than low) and by shifting consumption from one time to another (from high-priced to low-priced periods).

Because electricity prices are so temporally volatile, the traditional approaches used to assess the cost effectiveness of various electricity supply and demand options are no longer meaningful. The typical integrated-resource planning method compares the costs of an energy option to the annual average cost of electricity (either the dollar savings associated with energy reductions or the dollar payments for energy production). In the future, such analyses will have to consider temporal variations in output (consumption) and prices.

Exhibit 3. Real-Time Pricing and Residential Customers

Conventional wisdom holds that only the larger industrial and commercial customers will choose to face real-time prices, while other customers will continue to face time-invariant prices. While this situation will likely prevail during the early years of retail competition, we think many residential customers will elect to face time-varying prices in the long run.

A recent experiment with time-of-use (TOU) rates in Laredo, Texas, found high customer satisfaction with the program and substantial load shifts, as shown in the figure below. When prices reached their maximum of $350/MWh at 5 pm, participants cut their load by 36%. Averaged over the five high-priced hours, demand was cut 15%.

The current price for the customer equipment is less than $500 and is likely to decline due to technological advances.

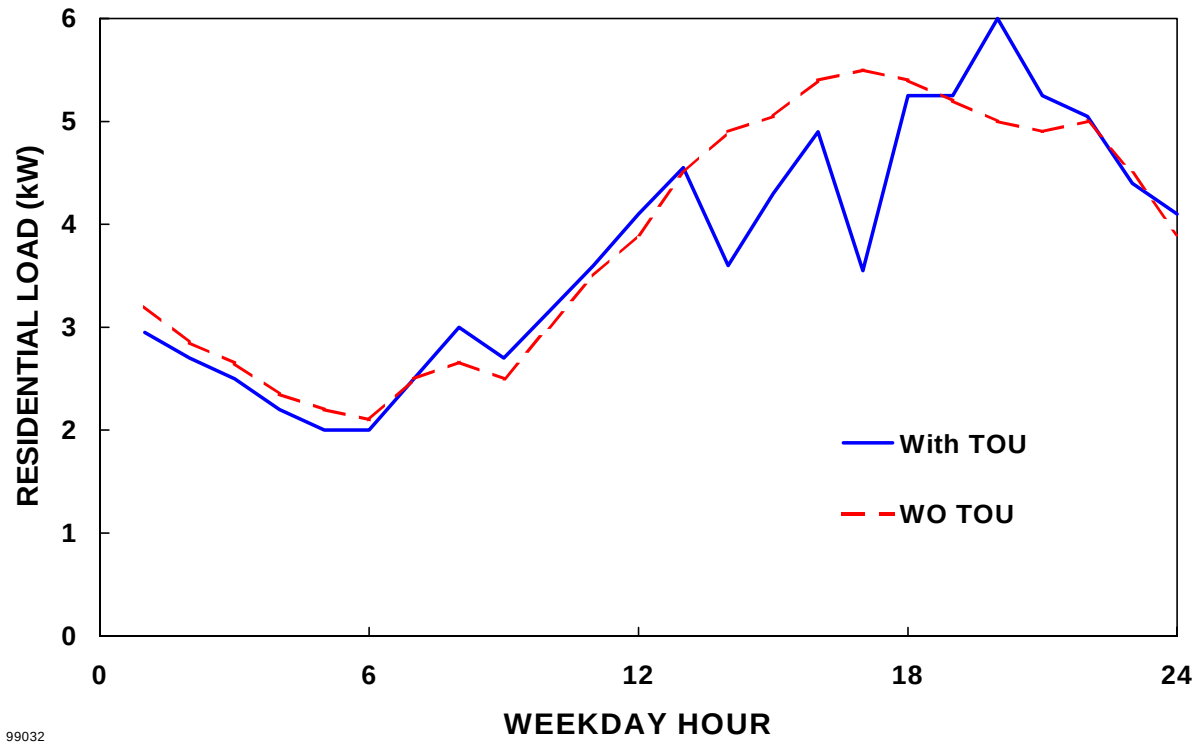


Figure 5, based on the day-ahead hourly prices in California, vividly illustrates the importance of time in assessing the benefits of any electricity-saving or -producing option. Averaged over the more than 13,000 hours during this 18-month period, the price of electricity was $25.7/MWh. Monthly prices ranged from $12 to $40/MWh, with a standard deviation of

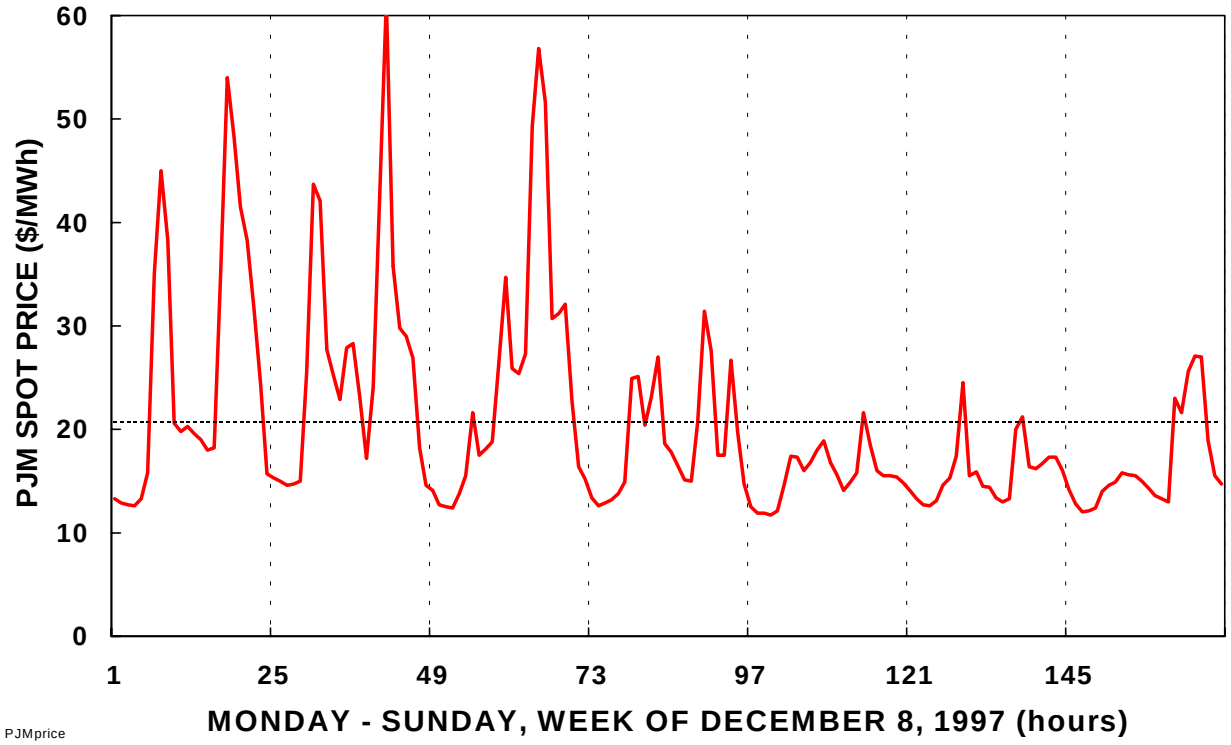


Fig. 4. Hourly prices in the Pennsylvania-New Jersey-Maryland (PJM) Interconnection for a week in December 1997.

$8.8/MWh. Weekly prices ranged from $7 to $52/MWh, with a standard deviation of $9.3/MWh. Daily prices ranged from $3 to $91/MWh, with a standard deviation of $11.8/MWh. Finally, hourly prices ranged from $0 to $225/MWh, with a standard deviation of $17.1/MWh. Figure 6 shows how hourly prices varied during this 18-month period; prices were less than $10/MWh for 9% of the hours and greater than $100/MWh for 1% of the hours.

Table 2 shows the economic benefits of various energy options as a function of their temporal flexibility. An option that either saves or produces electricity at the same level throughout the year has a benefit equal to the yearly average price of $25.7/MWh. An option that follows the actual load shape is 9% more valuable per kilowatt-hour. An option that saves or produces energy only during the months of August and September is 43 to 57% more valuable than one that operates at the same level hour-by-hour throughout the year. A resource that produces or saves energy only during those hours with the highest prices would be even more valuable.

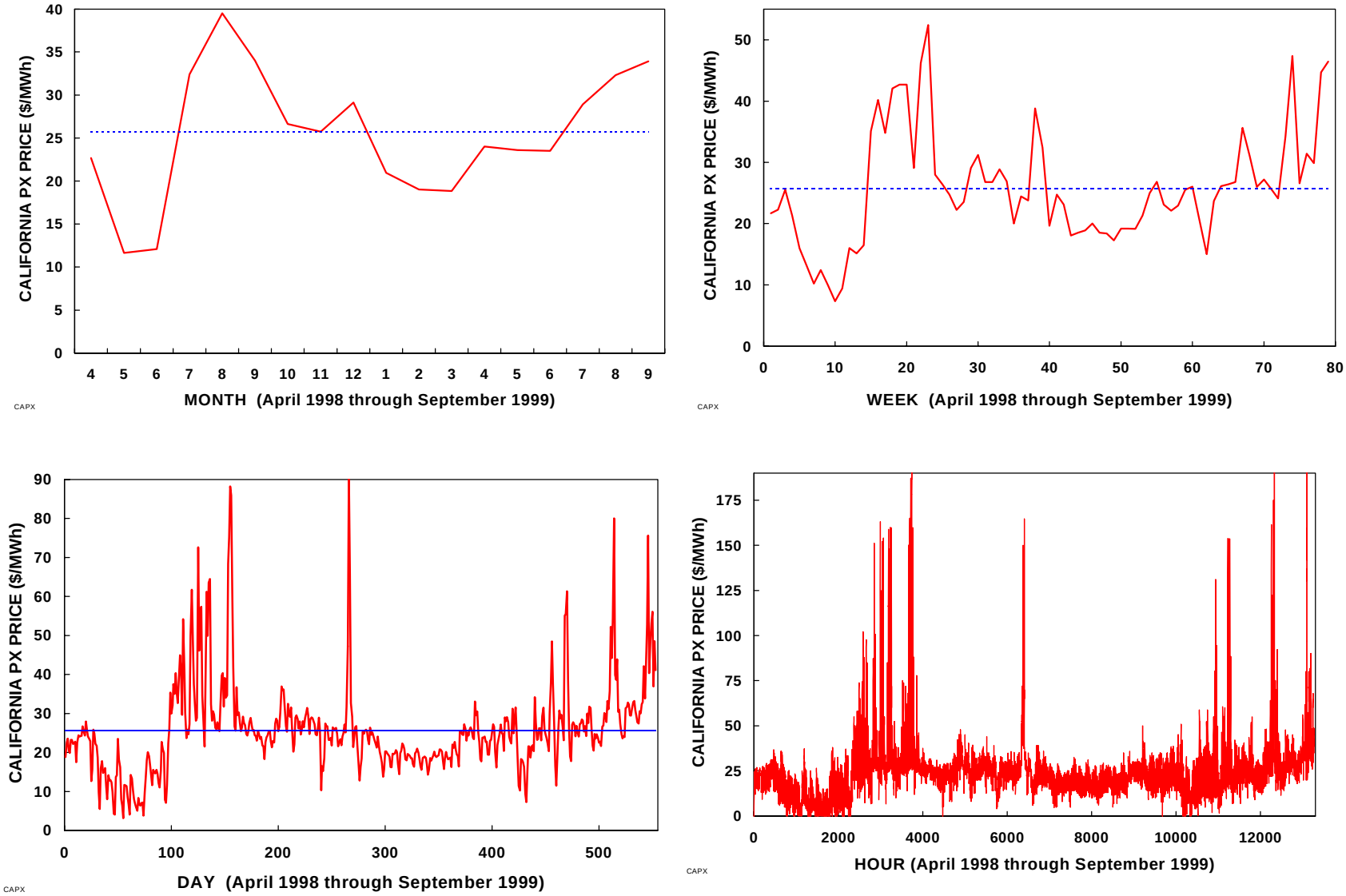


Fig. 5. **Monthly, weekly, daily, and hourly prices for electricity at the California Power Exchange from April 1998 through September 1999.**

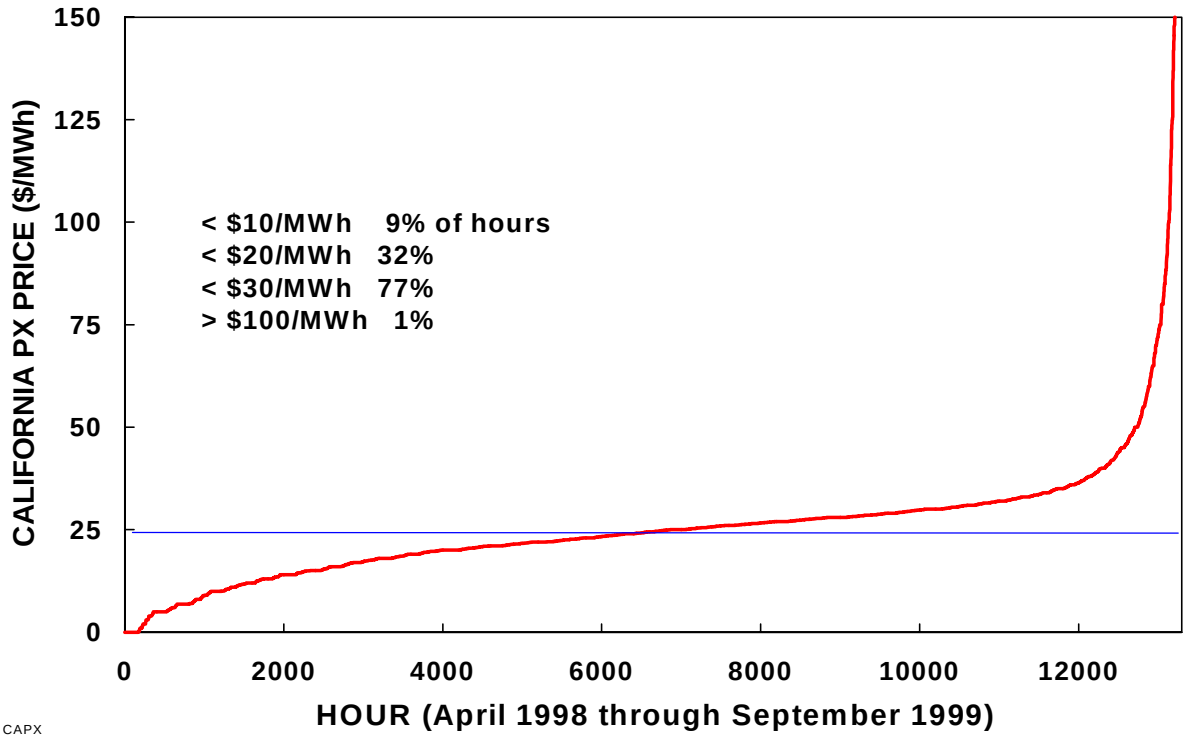


Fig. 6. **Price-duration curve for California day-ahead hourly electricity prices.**

Table 2. **Benefits of various energy supply and demand options relative to California day-ahead hourly prices**

	Benefit ($/MWh)	Ratio of benefit to 100%-capacity-factor benefit
Baseload resource, 100% capacity factor	25.7	1.00
Baseload resource, actual (60%) capacity factor[a]	28.0	1.09
Summer resource, 100% capacity factor	36.8	1.43
Summer resource, actual (68%) capacity factor[a]	40.4	1.57
Peaking resource, 20% hours with highest prices	46.4	1.81
Peaking resource, 10% hours with highest prices	59.7	2.32

[a]These are the actual load factors for California for either the full 18-month period or for August and September 1998.

The extent to which a resource can adjust its output (or consumption) in response to changes in electricity prices depends strongly on the technology. Some generators, such as nuclear units, operate at full output year round; they are taken offline only for refueling or

repairs. On the other hand, combustion turbines are operated only when prices are high. Their startup and shutdown times and costs are relatively low, which makes it possible for them to operate flexibly. Certain renewable resources, such as wind power, operate only when the "fuel" source is available (e.g., the wind blows or the sun shines) and are therefore nondispatchable and intermittent. Their ability to follow prices is limited. Demand-side resources display the same kind of range in flexibility (Exhibit 4). A program that improves the energy efficiency of water heaters is likely to save roughly the same amount of energy each hour of the year. A comparable program that focuses on air conditioners, however, is likely to be much more valuable per kilowatt-hour of savings because the price of electricity is much higher in the summer.

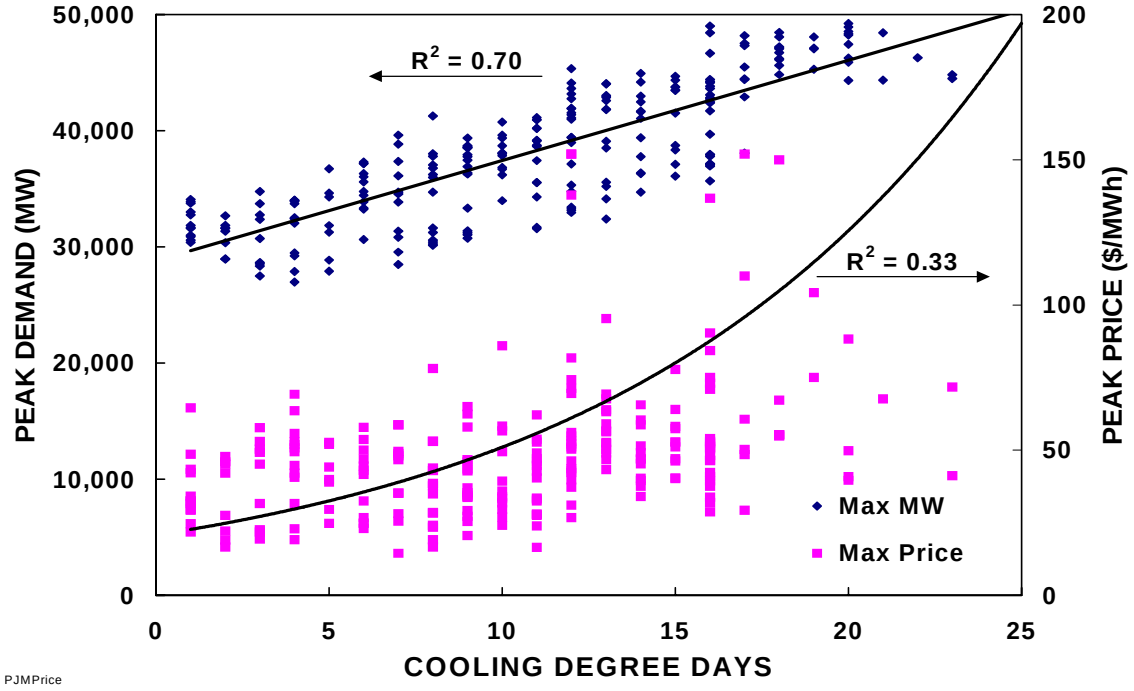
The situation that occurred in PJM on July 6, 1999, illustrates well the importance of timing. PJM's load reached an all-time high that day and, as a consequence, it deployed its active load management program to reduce demand during the mid-day hours (Fig. 7). The program cut demand by an average of 1% during nine peak-load hours. Because electricity prices reached $920/MWh during these hours, this demand reduction cut electricity costs by $10 million. Interestingly, the same electricity savings on the following day would have saved only $0.4 million because electricity prices on July 7 reached only to $50/MWh.

This example illustrates the likely differences between the effects and effectiveness of traditional utility demand-side management programs and the response of loads to market signals. Because real-time pricing was not available, traditional programs provided fewer direct benefits to participants. Although some of those programs provided important energy and environmental benefits, reliance on customer response to visible and volatile real-time prices is likely to have much larger effects on customer load shapes, although possibly smaller effects on overall load levels.

Exhibit 4. Loads, Weather, and Real-Time Pricing

Loads are very weather (especially temperature) sensitive, and prices vary with loads. As a consequence, demand-side measures related to heating and air conditioning are well suited to respond to time-varying electricity prices. Consider data on maximum hourly load and price for each day with positive cooling degree days in the PJM area (a total of 251 days between April 1998 and September 1999). Changes in maximum load alone explain almost 30% of the variation in maximum prices. The graph below shows (1) the relationship between maximum daily demand and cooling degree days (CDD), and (2) the relationship between maximum daily price and cooling degree days. Cooling degree days explain 70% of the variation in maximum daily loads. In addition, CDD explain 33% of the variation in maximum daily prices.

This simple example makes two points. First, summer (and, to a lesser extent, winter) temperatures have substantial effects on electricity demand and, therefore, on prices. As a consequence, measures that reduce customer demand at times of extreme temperatures are likely to substantially cut electricity bills. Second, much of the variation in electricity prices is *not* explained by weather data. Thus, efforts to respond to high spot prices require measures that go beyond automatic adjustments for temperature changes.



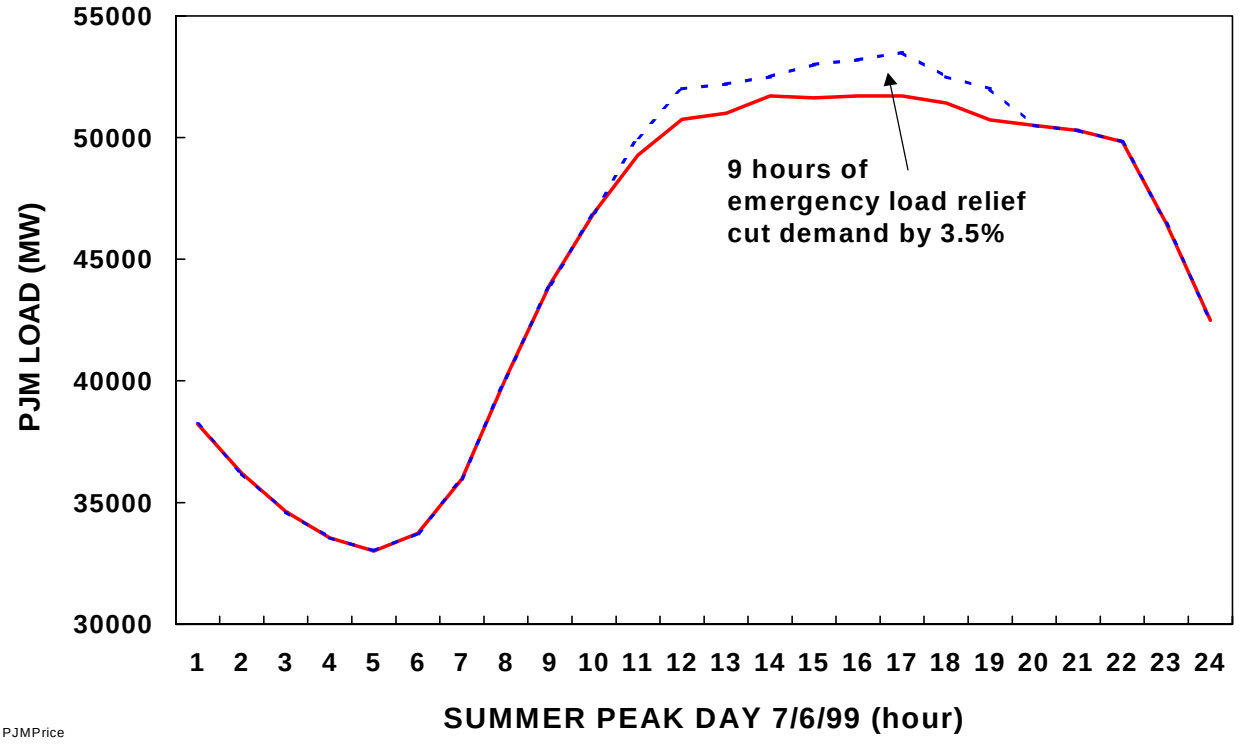


Fig. 7. **PJM hourly loads on its peak day in 1999, with and without the effects of active load management. Absent load management, peak demand would have reached 53,500 MW, 3.5% more than it actually did.**

INTRAHOUR PRICES

The importance of temporal variations in electricity prices is even greater than the discussion so far suggests. Prices vary not only from hour to hour, but also within each hour. The California ISO operates an intrahour balancing market that dispatches generators and sets prices every ten minutes. The PJM Interconnection sets prices every five minutes. The average PJM hourly price for August and September 1999 was $26.8/MWh. The maximum intrahour price difference during this period averaged $8.6/MWh, one third of the average.* The intrahour price difference was zero for only 5% of the hours, while the price difference exceeded $10/MWh for 20% of the hours in August and September.

Figure 8 shows how these intrahour prices varied for 12 hours on one weekday in September 1999. Intrahour prices were relatively constant for several hours, including hours ending 1, 3, 4, and 5 am. However, during some hours, especially from 6 am to noon, these prices varied substantially. Between 6 and 7 am, interval prices increased from $17 to

*The maximum intrahour price difference is the difference between the highest and lowest of the 12 5-minute interval prices within each hour.

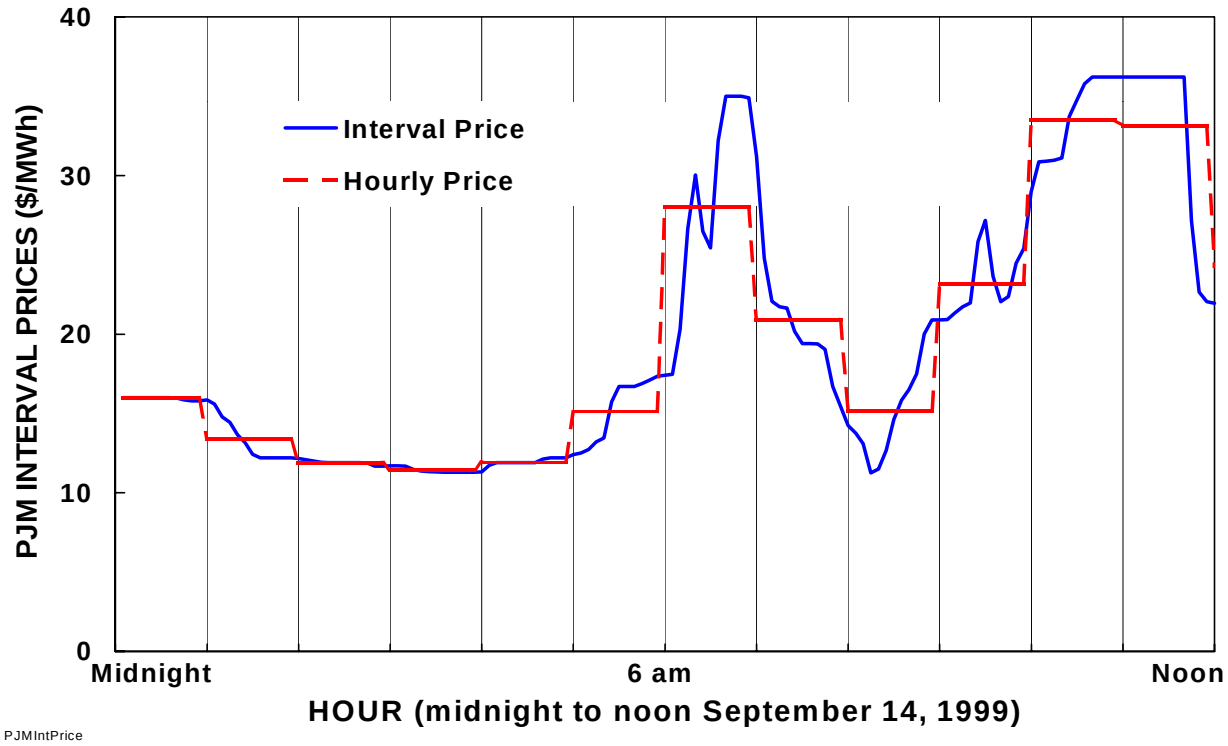


Fig. 8. Intrahour (interval) and hourly electricity prices within PJM for 12 hours on September 14, 1999. PJM calculates and posts prices every five minutes.

$30/MWh, then decreased to $25/MWh, and then increased again to $35/MWh. Whether a resource can capture the effects of these price swings depends on its ramp and acceleration rates, its minimum and maximum loads, its minimum run time, its startup times and costs, and the speed with which these price signals are communicated from the control center to the resource.

5. TRANSMISSION CONGESTION AND LOCATIONAL PRICES

Flows within a transmission system depend on the configuration of the system, the generation injections, and the load withdrawals. There is little ability to control flow within the network other than by removing elements (taking transmission lines out of service) or by changing the generation injections.*

Transmission congestion refers to the situation in which it is not possible to complete all the proposed transactions to move power from one location to another on the grid. Such commercial-transaction restrictions can arise because of thermal, voltage, or stability limits on transmission elements (see Exhibit 5). Congestion is generally *not* related to the actual flows

*Phase-angle regulators, DC lines, and flexible AC transmission systems (FACTS devices) provide limited, and expensive, control of flows at a few specific locations.

on lines. Congestion occurs when security-constrained dispatch requires modification of the economic dispatch. This situation occurs most frequently as the result of contingency analysis rather than because of steady-state line flows. The generation dispatch is modified because a line *will* overload *if* a specific contingency occurs (e.g., a generator or transmission line trips). Because there is often no time to take corrective action to prevent cascading failures if such a contingency occurs, it is necessary to preemptively modify the generation dispatch. It is this off-economic dispatch that results in locational price differences. (Losses also cause locational price differences but have a much smaller impact and are easier to deal with than congestion and are ignored here.)

Why is congestion so much more important a problem now than it was a few years ago? The traditional vertically integrated utilities accounted for transmission constraints when they made their daily operating (unit-commitment) plans. Thus, they used their generating resources in ways that would not overload the network. In today's increasingly competitive environment, suppliers schedule resources without a detailed knowledge of or interest in transmission constraints.

Exhibit 5. Transmission Limits

The maximum flows through transmission systems are governed by thermal, voltage, and stability limits:

- As the power flow through a transmission element increases, so too does the current. Heat is generated within the element proportional to the square of the current. The *thermal* limit represents the maximum current flow through the element, beyond which it will overheat and then sag, melt, or damage the insulation.

- *Voltage* limits refer to the reactive-power requirements of transmission elements. Here, too, as the power flow through a transmission line increases, so does the amount of reactive-power needed to maintain voltages. If the reactive resources are insufficient voltages will decay and may suddenly collapse, leading to an interruption.

- Finally, *stability* limits relate to the requirement that all generators in an Interconnection operate at the same frequency and phase angle. If a disturbance causes a disequilibrium that is not quickly restored, the system can come apart causing major outages.

In the long term, construction of new generators and transmission lines can reduce congestion. In the short term, system operators can treat congestion in two basic ways. They can mandate engineering solutions or they can use prices to let suppliers and consumers (i.e., market participants) decide which transactions to cut. As shown by the simple example in

Fig. 9, the costs and effects of the two approaches can be quite different. In this example, cutting Generator 2 by 60 MW costs only $300 [60 × (25 – 20)], $200 less than the cost of cutting Generator 1 by 50 MW [$500 = 50 × (25 – 15)].

The simplest (engineering) approach is to ignore congestion in setting energy prices (i.e., assume that all proposed transactions can be completed as if the amount of transmission capacity was infinite). If proposed transactions threaten to overload transmission lines, the security coordinator implements NERC's transmission loading relief (TLR) procedure. This procedure adopts an engineering approach to congestion relief. Transactions that contribute 5% or more to the congestion are curtailed depending on their firmness, with nonfirm transactions cut before firm transactions are cut. Many market participants oppose TLR because they believe that the incumbent utilities manipulate the TLR calls and implementation to favor their own transactions. In addition, FERC opposes the current TLR procedure because it is economically inefficient.

An alternative approach is to socialize congestion costs. With this approach, the system operator pays generators on either side of the constraint to increase output (constrained on) or decrease output (constrained off) to relieve the congestion. The system operator pays these generators for any opportunity or out-of-pocket costs associated with this uneconomic dispatch. The costs so incurred are then allocated to all transmission customers through an uplift charge. Although simple to implement, this approach is economically inefficient because it fails to send price signals to transmission users on the true costs associated with their transactions. The absence of location-specific prices also robs investors of important information on where to locate new generators and what transmission projects to build.

The economically efficient way to deal with congestion is to use locational prices that signal transmission users on the actual costs of transmission service. Locational prices can be set on a zonal basis (as in California) or on a nodal basis (as in PJM). New York and New England plan to use a mixed system, in which generators will face nodal prices and consumers will face zonal prices.

PJM calculates prices within its control area for about 2,000 nodes. Between April 1998 and September 1999, the average hourly price in PJM was $27.4/MWh. During this 18-month period, prices differed from location to location for 15% of the hours. During these congested hours, the maximum locational price difference averaged $19/MWh. These locational differences vary from hour to hour and from month to month. During 1998 and 1999, both the fraction of hours with nonzero locational differences and the magnitude of these differences were much greater in May, June, and July than in January and February.

These data on locational price differences suggest that the benefits of various energy supply and demand options depend not only on their temporal flexibility but also on their location. Unfortunately, the locational picture is complicated because the direction of congestion sometimes changes. For example, within PJM the predominant flows are from the

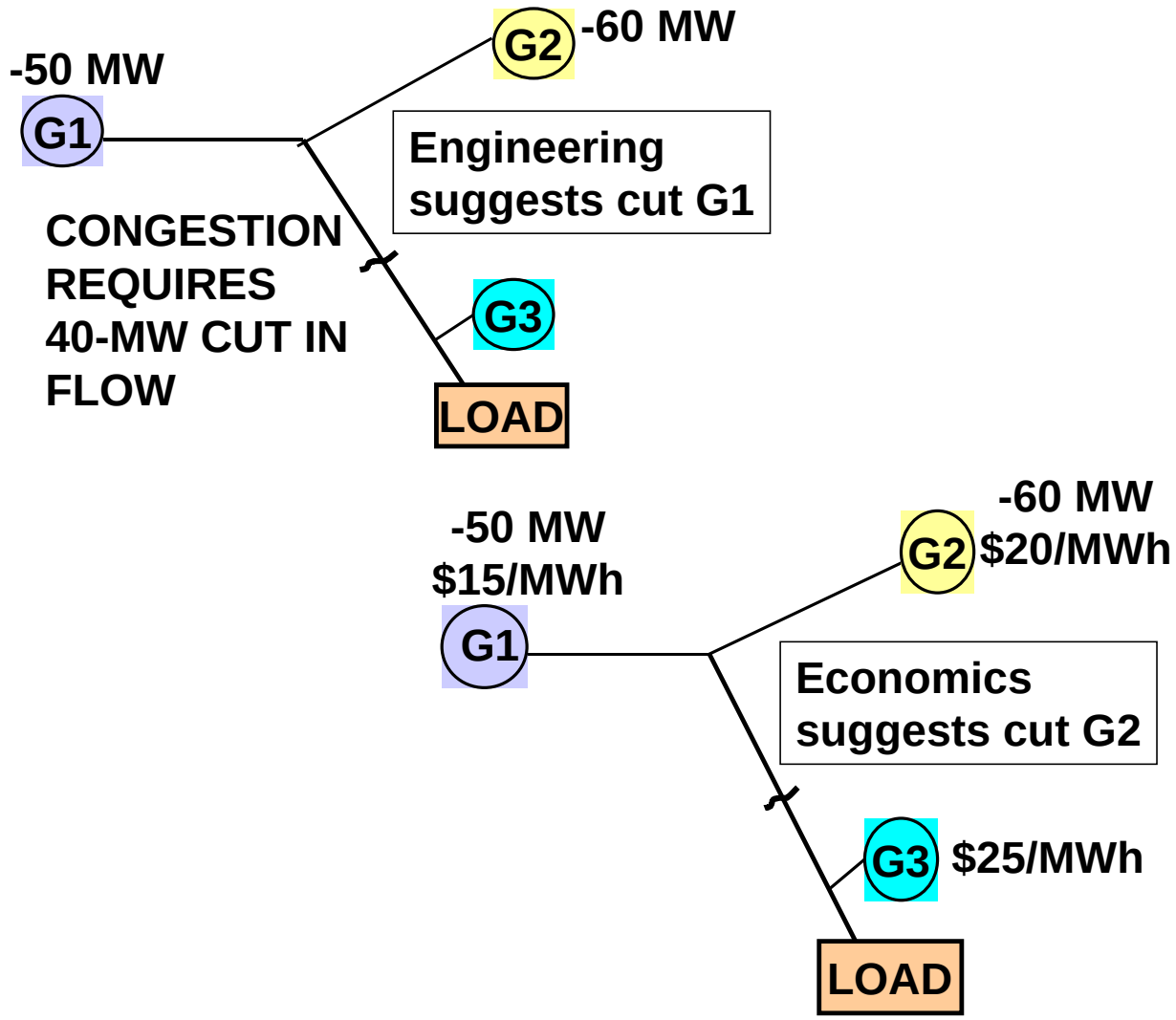


Fig. 9. **Engineering (left) vs economic (right) methods to relieve transmission congestion. Even though relieving the congestion by cutting output at G1 has less physical effect (fewer MW cut), the economic cost is greater than cutting output at G2.**

west (where the inexpensive generation is located) to the east (where most of the loads are located). However, the flows sometimes change directions. During the summer of 1999, for example, PJM prices were higher in the east 20% of the hours, the same throughout the region 75% of the hours, and higher in the west 5% of the hours.*

*The same phenomenon occurs in California, where the primary congestion is between the north and south. When the Pacific Northwest has ample supplies, flows are primarily from north to south, with prices higher in the south. However, during offpeak periods, flows are often from south to north, with prices higher in the north.

6. RESOURCE PLANNING AND INVESTMENT

Previous sections dealt with the reliability and commercial aspects of bulk-power system operations. In this chapter, we focus on what NERC calls adequacy. Although adequacy is a reliability concept, it has strong commercial implications.

Obviously, adequacy and security are complements. Without system security, the output of the generation resources, no matter how abundant, cannot be delivered to customers. Correspondingly, a high degree of security is of little value if there are insufficient resources to meet customer needs.

Adequacy and security can also be substitutes; more of one can make up for less of the other. For example, an abundance of resources makes it easier to maintain a high degree of security (i.e., reduces the need for emergency actions). That is, system operators can manage the system in real time with less data and fewer analytical tools if there are ample generation resources and redundant transmission facilities. Similarly, high-quality system operation can extract more output from a system that might otherwise be considered underbuilt. For example, the near-real-time collection and analysis of data on the current and projected states of the transmission system can allow system operators to run the system closer to its limits than would be feasible with less data collection and analysis.

A restructured, competitive electricity industry will reduce the integration between generation and transmission planning. The Energy Modeling Forum emphasized the close coupling between generation and transmission:

> The daily operation of the transmission system depends critically upon where and when to generate power. Longer-run decisions about investing in generation or loads are closely linked to those concerned with expanding the transmission system. The existence of these interrelationships, or complementarities, between functions presents opportunities to operate and expand both systems more efficiently or at a lower cost when done jointly rather than separately. A fundamental issue in restructuring concerns how to decentralize decisions about generation and loads and still acknowledge the complementarities between generation and transmission.

Historically, it took up to a decade or more for utilities to plan, gain regulatory approval for, and build new generating units and new transmission lines. Public and environmental concerns have, for at least a decade, complicated and lengthened the transmission-planning and -expansion process. The transmission-planning processes being established by ISOs are inherently slow because (1) they are serial (i.e., transmission is planned only after new generation facilities are announced), (2) they encourage participation by all stakeholders, and

(3) they require cooperation and agreement between the ISO and the transmission owners.* This slowdown in transmission construction is occurring at the same time that new gas-fired technologies and competitive generation markets are shortening the time it takes to plan and build new generation.

DATA AND PROJECTIONS

The U.S. electricity industry is currently in a very awkward position—half regulated and half competitive. Many utilities are understandably reluctant to make investments until the rules and the separation between competitive and regulated activities are clear.

Figure 10 based on data reported annually by U.S. investor-owned utilities, shows that transmission investment relative to total energy production declined slowly between 1988 and 1997. Transmission maintenance relative to energy production declined more rapidly during this time. It is unclear whether these declines represent a degradation in transmission-system reliability or improved productivity.

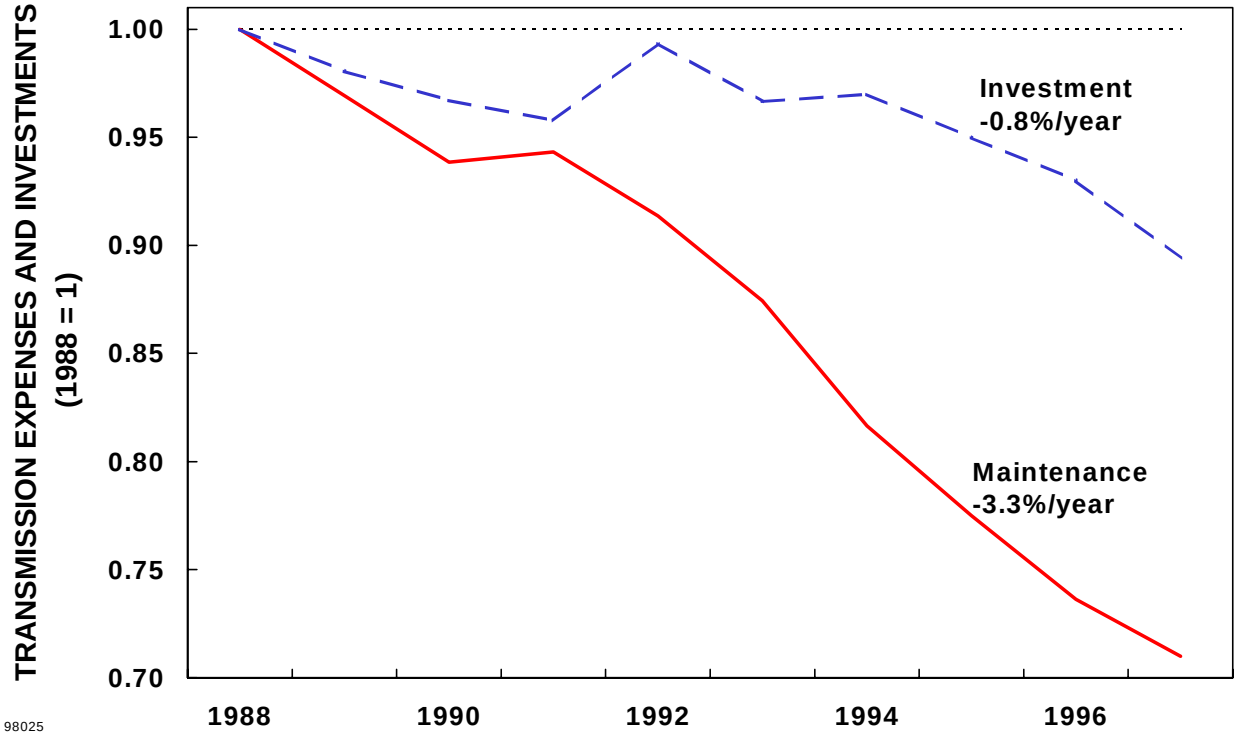


Fig. 10. Trends in annual transmission maintenance expenses and investment for U.S. investor-owned utilities, normalized by annual electricity production.

*Because Transcos both own and operate transmission, this factor does not apply to them.

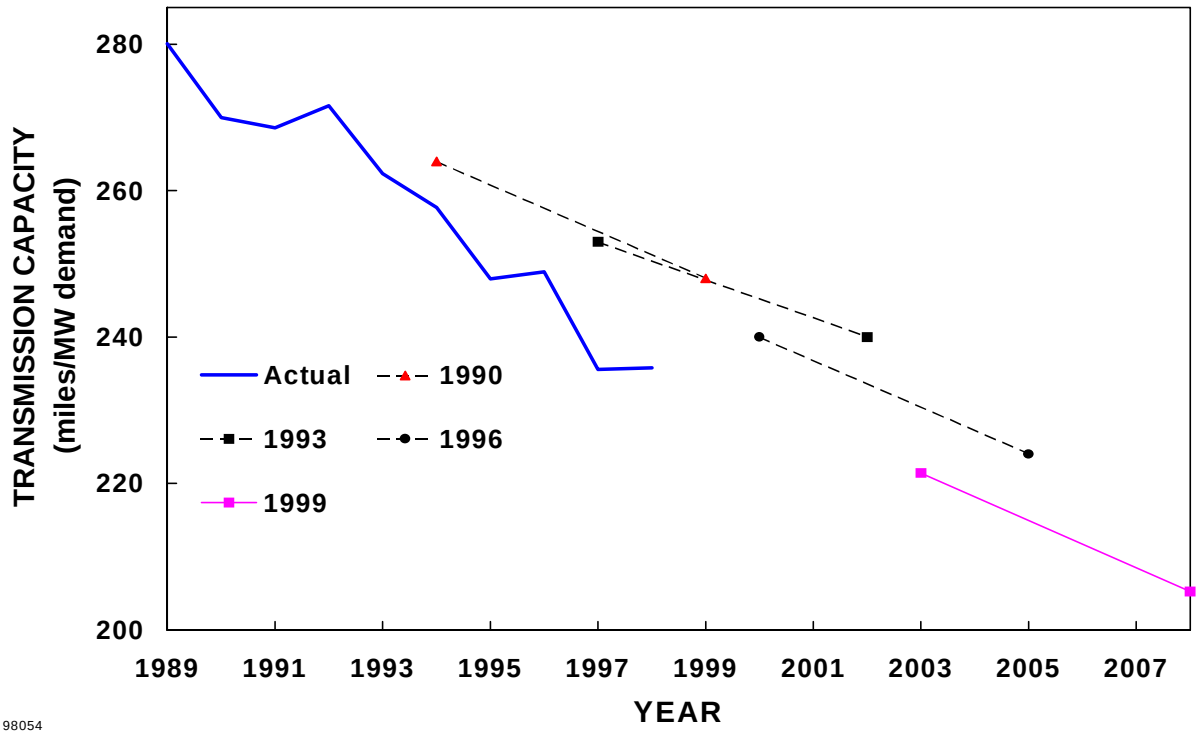


Fig. 11. U.S. transmission capacity per MW of peak summer demand from 1989 through 1998 and projections to 2008.

Data reported to NERC show similar trends: a 16% decline in miles of transmission lines per MW of summer demand between 1989 and 1998 (Fig. 11). Transmission capacity is expected to decline an additional 13% by 2008.

Figure 12 shows utility forecasts of generation-capacity margins from 1990 through 1998 and projections through 2008. Nationwide, reserve margins declined from 22% in 1990 to 16% in 1998 and are expected to decline further to 9% in 2008. As NERC notes, the projections for the last five years are highly uncertain. This uncertainty occurs because the owners of merchant plants often do not reveal their plans early and because new generating units can often be constructed in only a few years (reducing the need for long-term projections of generating capacity). Estimates from the Electric Power Supply Association paint a more positive picture of future generation adequacy than do the utility estimates; EPSA data suggest that up to 99,000 MW of new capacity will be built by the year 2003.

GENERATION ADEQUACY

Historically, utilities maintained “extra” generating resources for short- and long-term purposes; this section focuses on long-term reserves, often called planning reserves, and does not deal with operating reserves. At least two mechanisms can be used to maintain generation adequacy:

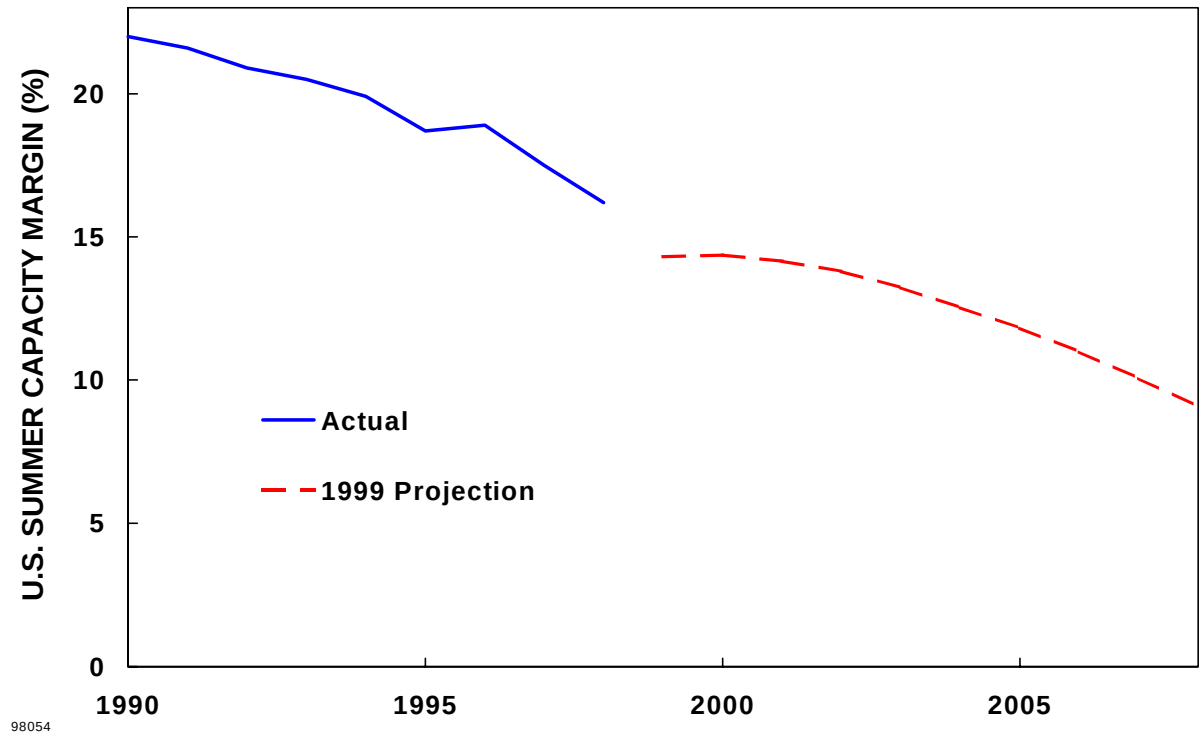


Fig. 12. U.S. summer generation-capacity margins from 1990 through 2008.

- Rely on markets, the interactions of consumers and suppliers acting through the mechanism of volatile spot prices, to decide what types of generation to build and when and how much electricity to consume when. California adopted this approach.

- Rely on the traditional system of having a central agency [e.g., the RTO or state regulator] specify an appropriate minimum reserve margin based on estimates of the value of lost load (VOLL) and other factors (e.g., forced and planned outage rates for different types of generating units). This reserve margin is then imposed on all load-serving entities (LSEs). The three Northeastern ISOs (PJM, New York, and New England), all of which developed from traditional tight power pools, use this approach.

These two options should produce different outcomes in:

- hourly energy prices, with reliance on real-time markets likely to yield lower average prices and costs but greater price volatility;

- customer load shapes, with reliance on real-time markets likely to yield higher load factors; and

- generation portfolios, with reliance on real-time markets likely to yield relatively more baseload capacity.

Most market participants believe that—in the long run—generation adequacy will be left to markets with little involvement by government regulators. To do otherwise would interfere with the workings of competitive energy markets. That is, the energy and capacity markets are closely coupled.

On the other hand, we may need a multiyear transition period while suppliers and, especially, retail customers learn how to respond appropriately to rapidly changing (e.g., hourly) electricity prices. We first have to permit retail customers to face these time-varying prices; in most parts of the country, customers still face prices based on embedded costs that are largely time invariant. We also need to establish intrahour (real time) balancing markets, as required by FERC in its Order 2000 on RTOs. During this transition period, prudence may require maintenance of mandated planning-reserve margins.

Figure 13 schematically illustrates the supply/demand balances with and without an explicit installed-capacity requirement. The dashed line that slopes up to the left represents consumer demand, and the stairstep line that slopes up to the right represents generating capacity. With a reserve-margin requirement of 11,500 MW, supply and demand equilibrate

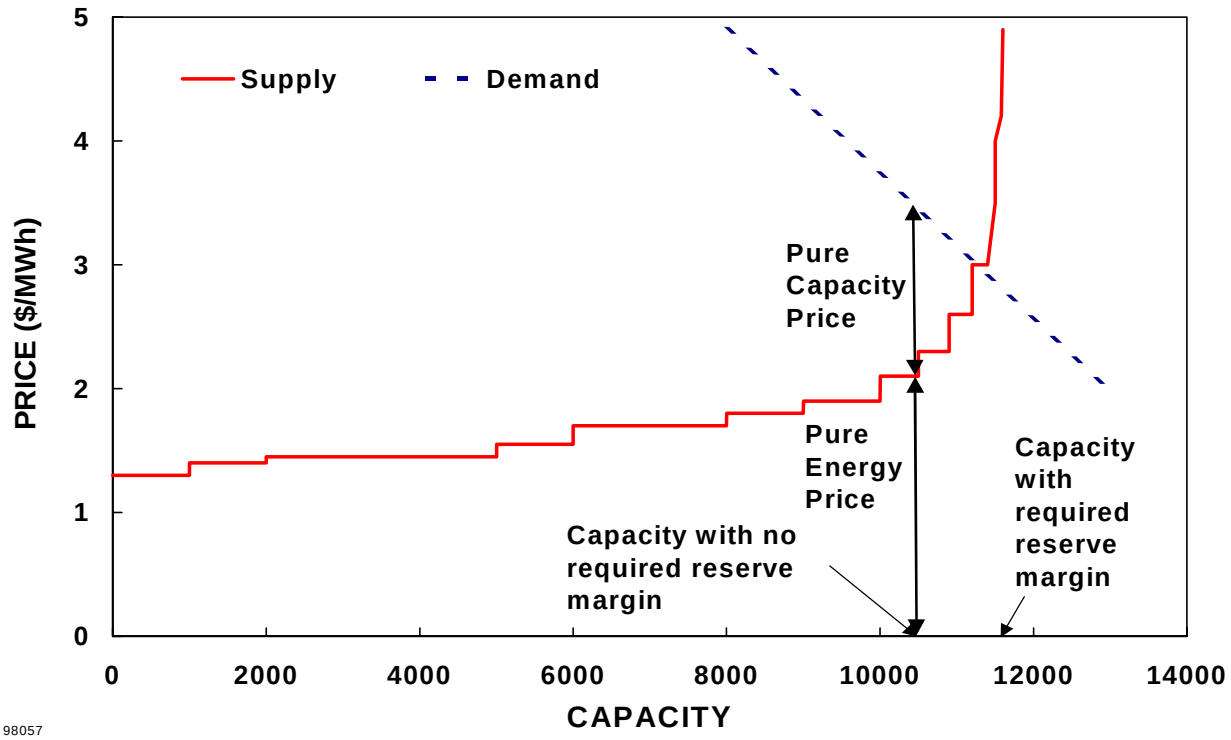


Fig. 13. **With an installed-capacity requirement of 11,500 MW, supply and demand balance at a price of 3.0¢/kWh. With no capacity requirement and only 10,500 MW online, unconstrained demand would exceed supply, and prices would rise to 3.4¢/kWh.**

at a price equal to the variable cost (fuel plus variable operations and maintenance) of the last (marginal) unit online at that time. If, however, there is no required reserve margin and market forces yield only 10,500 MW of available capacity, the price of electricity will rise above the variable cost of the last unit online when unconstrained demand exceeds 10,500 MW. The amount of price increase (the pure capacity price in Fig. 13) is a function of the demand elasticity for electricity. The more responsive customer demand is to changing electricity prices, the smaller this capacity price will be.

This example makes two points. First, even if there is "insufficient" capacity from an engineering perspective, price-responsive demand and supply will equilibrate, and the bulk-power system will not crash. This equilibrium occurs because some customers would rather forego some consumption than pay the high price associated with this situation. Second, at times of high demand, spot prices will be higher if there is no required reserve margin. In other words, specifying a minimum amount of installed generating capacity will suppress spot prices at certain times. Economists argue that this suppression of a valuable price signal will undercut energy and capacity markets.

Requiring a minimum reserve margin creates two markets (installed capacity and energy) with no assurance that they will be in equilibrium. This requirement will suppress energy prices and undercut demand-side participation in reliability. On the other hand, energy-only markets will encourage efficient capacity planning, which is where the primary benefits of competitive generation lie, and encourage demand-side participation in reliability markets.

Reliance on market prices can work well only when spot prices accurately reflect costs. Ruff notes that, "In practice, hourly energy prices tend to be below the average of the 'correct' instantaneous prices over the hour, particularly during critical hours when peaking capacity is needed. This is because the ISO must use a relatively simple, mechanical rule to determine a single energy price for the hour ... and most simple pricing rules miss within-hour effects."

Jaffe and Felder believe that mandated capacity requirements are needed because such capacity benefits society at large, not just the owners of such capacity. Such societal benefits are especially large for electricity because of its pivotal role in modern society, the real-time nature of electricity production and consumption (which occur within milliseconds of each other), and the difficulty of storing electricity.[*] They note that policymakers can either set minimum-reserve margins or subsidize capacity with an up-front $/kW-year payment for capacity. In principle, the two approaches should yield the same outcome.

NERC raises concerns that "few, if any, customers understand the implications of contracting for other than firm power supplies and firm transmission services." Because of the

[*]These societal benefits might include avoidance of the looting and violence that can erupt during a major blackout and the maintenance of electrical service to vital societal functions, such as hospitals, police and fire stations, traffic lights, and airport traffic-control systems.

long tradition of ample supplies and the use of interruptible rates to offer implicit discounts to large industrial customers, these customers are used to very few interruptions in service. Indeed, industrial customers, when interrupted, often are angry. Thus, it is an open question how customers will respond to real-time pricing. In addition, only a few electric utilities (e.g., Georgia Power) have much experience and a clear understanding of whether and how customers might respond to real-time pricing.

TRANSMISSION ADEQUACY

Transmission adequacy poses tougher problems than does generation adequacy for at least four reasons. First, system operators cannot easily control power flows over individual transmission elements, undercutting the notion of the traditional contract path and its associated transmission prices. Second, flows on one transmission element affect flows elsewhere on the grid, which creates loop flows and attendant problems in other parts of the grid. Third, transmission construction involves large economies of scale and scope, which means it is much cheaper to build more transmission at one time than is needed for now and that the benefits of such construction are widespread. Finally, transmission costs are almost all capital rather than operating, which means it is difficult to design economically efficient rates that also recover transmission costs.

The basic function of transmission is to interconnect loads and generators. Originally, this function was accomplished by vertically integrated utilities to expand supply options and improve reliability. Utilities discovered that the same interconnections designed to share reserves were also useful for economic transactions. Of course, even the use of transmission to enhance reliability was driven by economics; it was cheaper to share reserves than to build redundant generation-reserve capability.

The traditional planning process balanced generation to load but had little need to facilitate competitive generation markets. A vertically integrated utility strives to have sufficient transmission to *economically* supply its load. Power system planners forecast load patterns and generation-resource availability. Historical performance, including the load at each bus, was used to create detailed models of the electric system for peak and off-peak conditions. Models, data, analyses, and transmission plans were coordinated with neighboring systems.

The California ISO lists six reasons that transmission enhancements may be required:

- Interconnect generation or load (e.g., build a radial line from a new generator or load to the transmission system)
- Protect or enhance system reliability (e.g., replace older, less reliable equipment with newer, more reliable equipment)
- Improve system efficiency (e.g., replace high-loss equipment with lower-loss equipment)
- Enhance operating flexibility (e.g., add switching capability)

- Reduce or eliminate congestion (e.g., add new transmission lines or increase the capacity of existing lines)
- Minimize the need for must-run contracts (e.g., add transmission lines or reactive support at locations that depend on a single generator)

It is difficult to separate reliability from commerce in determining the motivation for a particular transmission project. Many real-world enhancements address multiple needs. An additional line bridging a congested interface would probably reduce congestion, increase reliability, and improve efficiency. It might also increase operating flexibility and minimize the need for must-run contracts.

A basic question should be addressed at the outset. Because transmission costs only one tenth as much as generation, why not build enough transmission so that it never constrains generation markets? This approach is appealing to engineers who like systems with flexibility and sufficient excess capacity to meet unexpected needs and to market participants who want to buy and sell power over large geographic regions. However, the construction of new transmission lines is often opposed by local residents and landowners and is therefore politically difficult to achieve. Building enough new transmission facilities to eliminate congestion is infeasible because congestion depends on the locations of generation and load, not just their magnitudes. Finally, regulatory rules may not permit utilities to recover fully the costs of such "overbuilt" systems. Who should pay these costs is also unclear.

Planners test system adequacy by modeling performance (line flows and bus voltages) under a full range of expected load, generation, and contingency conditions. Limits on the acceptable generation dispatch range are determined for each set of operating conditions. The planners then make a judgement as to the adequacy of the transmission system.

Sometimes the judgement is straightforward. When the load being served by a radial line exceeds the line's capability, the transmission system must be enhanced, or local generation must be added. The need for enhancement is less clear in the more common case when inadequate transmission results in constraints on economic dispatch of generators rather than forcing curtailment of load. The complexity is twofold. First, the increased cost that will result from constraining the dispatch is generally an operating cost whose magnitude depends on the number of hours a year the constraint exists and the relative costs of the generators involved. At the same time, the transmission-enhancement cost is primarily a capital cost, which must be recovered over several decades. Which solution is the more economical depends on a number of factors, such as fuel costs, the cost of capital, and the expected locations and magnitudes of load growth and generation construction. Second, the need is less clearly reliability based. There may be opposition to using eminent domain to address an economic problem, especially if much of the economic benefit accrues to electricity consumers or providers in another state.

Restructuring further complicates this situation. Organizationally separating transmission from generation means that transmission planners must forecast the commercial

actions of generation owners (i.e., the timing, location, and size of new generating units) as well as load growth. The generation forecast must account for the operation of generators under market conditions rather than simply the cost differences that were considered when utility planners optimized over generation *and* transmission. The effects of these decisions are different as well. A generator that is constrained out of the market for too many hours a year may be driven out of business because it cannot recover its fixed costs. If market conditions change such that a transmission project is no longer economical, on the other hand, the cost is borne by the customers of this regulated asset.

Locational Pricing

Competitive markets for generation and retail services can be accommodated with monopoly management of transmission operations and investment. The monopoly could take on the obligation to provide transmission service for everyone. The monopoly would make investments and/or pay for redispatch to manage congestion. The National Grid Company in England and Wales works essentially this way; so might the various U.S. Transco proposals. All the usual problems with regulating and providing appropriate incentives to a powerful monopoly still exist.

Locational pricing is an alternative to monopoly management of transmission. In spite of its complexities, several schemes have been proposed for pricing transmission based on locational electricity prices. Hogan describes transmission-congestion contracts (TCCs) that are equivalent to tradeable physical transmission rights:[*] "With such contracts to allocate transmission benefits, it would be possible to rely more on market forces, partly if not completely, to drive transmission expansion." In Hogan's scheme, transmission service is priced partially on the nodal price differences. Nodal price differences result when transmission is congested and there is insufficient transmission capacity to move lower-cost power to higher-cost locations.

TCCs do not give the contract holder the right to move an amount of power over a particular line or between a particular pair of locations. Instead, they give the holder an income based on the price difference between the two locations. This achieves essentially the same result without disrupting the actual generation dispatch. To see how this works, consider the generator G and load L in Fig. 14. L contracts for 200 MW from G at $40/MWh.

G sells its output through the PX at the locational spot price. L purchases power through the PX at its locational spot price. The TCC operates to turn these two spot transactions into a firm power contract with guaranteed transmission. If the price at both L and G rises to $50/MWh, L pays $50/MWh, G receives $50/MWh, and G pays L $10/MWh. The net result is the same as L paying G $40/MWh. If congestion between G and L results in the price at L rising to $60/MWh, G's TCC results in a $10/MWh payment ($60 – $50) to G. L pays the PX

[*]Other proposals differ in detail but encompass the same basic requirements.

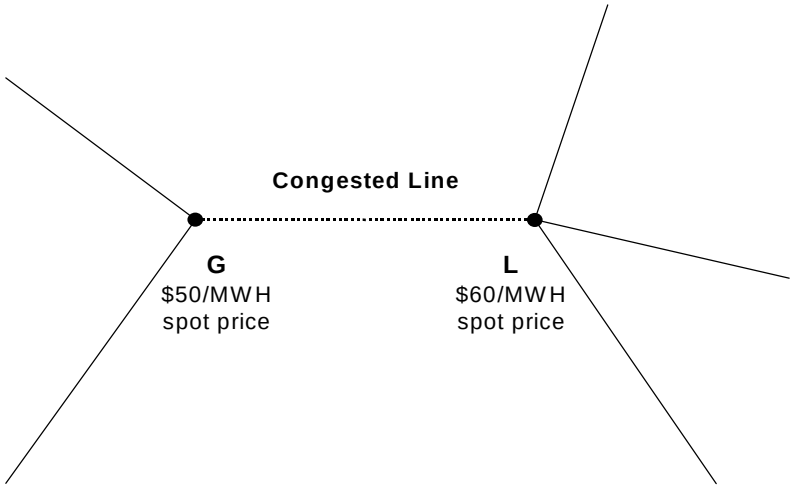


Fig. 14. TCCs are the economic equivalent of firm transmission contracts between locations with different power prices.

$60/MWh, G receives $50/MWh from the PX. G then pays L the $10/MWh TCC payment and the $10/MWh excess profit it received from the PX ($50 – $40). Again, the net result is a transaction between G and L at $40/MWh. The PX retains none of the TCC revenues.

TCCs could be awarded in various ways. Existing capacity could be divided to reflect historical rights. Alternatively, TCCs could be auctioned off. While the allocation process may be very important to individual market participants, only the fact that the TCCs are allocated is important to the system. Once allocated, TCCs allow markets to appropriately value and allocate scarce transmission among competing users.

In defining the TCCs, making sure they are simultaneously feasible is essential. This is done by modeling the transmission system. In all cases, the ISO or a neutral analysis party determines the transfer capability from each bus to every other bus. This transfer capacity is then auctioned or allocated, no matter where the constraint actually lies. The constraint's impact on the transfer capacity between each pair of buses is what is important.

The owner of a TCC has a specific MW of capacity between two buses. Whenever the spot price deviates at the two buses, the TCC owner gets the differential price times the TCC congestion capacity. The TCC owner may have no idea who is transacting to create the congestion. Only the system operator has sufficient information to administer the TCC payments.

Looking only at the locational marginal prices and the defined capacities between specific locations makes it possible for system users to ignore the physical details (the system

operator must remain keenly aware of the physical system). In fact, a market can be administered that converts the physical network into a hub-and-spoke system, a much simpler trading structure.

The system operator should not be allowed to keep the excess TCC payments (in general, there will be more congestion payments collected from locational price differentials than are paid out to congestion-contract owners). If congestion payments remained with the system operator, this would give it an incentive to create congestion. Instead, excess congestion payments should be used to offset the fixed costs of transmission.

Transmission Investment

TCCs can help allocate scarce transmission among competitive users in the short term and they can help with transmission investment in the long term. Congestion contracts covering the additional capacity created by transmission enhancements are awarded to the investor. The system operator uses the same system models that allocated the original TCCs to determine what additional TCC capacity is created by the transmission enhancement. All existing TCC rights are preserved, and the investor gets the new TCC rights.

Although the investor gets the new MW capacity, congestion prices are likely reduced, possibly to zero. So, although everyone's *rights* are preserved, the *value* of those rights can be wiped out. Hence, congestion contracts will most likely be tied to the generators or loads that profit from reduced congestion rather than to investors building transmission for profit. Although generators or loads lose the value of their congestion contracts, they gain more from reduced power prices at the receiving bus or increased prices at the sending bus.

This impact of transmission enhancement on locational prices creates a free-rider problem that may make it difficult to get investors to invest. Each party is better off if someone else makes the investment. The problem is compounded by the scope and scale problem. It would be desirable to invest only enough to cover the exact amount of transmission capacity required by the individual needs. Then anyone else will cause congestion and there will be a price differential that the investor's TCC will collect rent from. But it is generally much cheaper to oversize a line when constructing it than it is to come back later and upgrade it. Similarly, in an interconnected system, economies of scope often mean that an enhancement removes constraints on a number of transmission paths. Customers using one path may choose to wait until customers using another path make the investment. The longevity of transmission equipment (roughly 50 years or more) adds a further complication.

Even if all of the current beneficiaries could agree on the need for a transmission enhancement, they are unlikely to follow through. A user of the transmission facility would have to be able to commit for decades of benefits to pay for the investment. Since the enhancement immediately becomes a sunk cost, providing the benefits at nearly zero

incremental cost, there is no ability for the current user to sell its share in the investment if it no longer needs it unless the system becomes congested again.

In cases where it is not practical for specific users to invest to relieve congestion, it may be appropriate to rely on congestion pricing to identify economically sound investment. Once identified and approved by regulators, the transmission enhancement could be built, either by the existing transmission owner or by a third party. The enhancement could then be added to the rate base for all users of the system or for users of a portion of the system.

Pricing Principles

While congestion contracts may seem complex, several fundamental principles concerning pricing transmission in a restructured industry are generally accepted:

- No one can be allowed to withhold transmission rights; unused capacity is available for others to use.
- The system operator must analyze the rights and calculate the payments for transmission congestion and losses.
- The system operator must not be allowed to profit from congestion.
- Transmission prices should reflect differences in locational power prices.
- Transmission prices can send appropriate signals for the locations of new transmission and generation facilities as well as loads.
- Transmission prices can illuminate the need for transmission enhancement but they are unlikely to provide sufficient incentive to motivate investment without additional compensation.
- The remaining transmission costs will have to be allocated among users.

These principles, unfortunately, do not lead to a completely market-based expansion policy. This situation leaves most enhancement decisions in the discouraging position of having many interested parties and no uniquely correct solution. To make matters worse, multiple solutions often are, from the system's point of view, equally good (i.e., low in cost). But the differences can be critically important to individual market participants. With a vertically integrated utility, differences in the least-cost solutions did not matter. The customers only paid the aggregated cost. Now it is very important because individuals prosper or starve based on the final decision. For example, a generator located within a congested portion of the grid might be driven out of business if the congestion is relieved, a dramatic result for the owner of that unit. The amortized capital cost of relieving the congestion, on the other hand, may be only slightly lower than the off-economic dispatch cost, a small impact on the overall system. The fact that the decision to invest will, of necessity, be based on forecasts makes the problem worse. Finally, the analysis and decision will likely be made by entities (the ISO and the regulator) that bear no market risk themselves.

Transmission-Expansion Approaches

As regional transmission entities develop in different ways throughout the country, they are adopting and applying different transmission planning and investment decisions. To date, these entities have not developed market-based methods to decide on transmission expansions; explicit links between congestion pricing and transmission investment have not yet been made. Entergy, FirstEnergy, and the Alliance RTO recently applied to FERC to establish Transcos; however, none of the filings says much about transmission planning and expansion.[*] All three note the benefits of combining transmission ownership and operation in one entity. Entergy states that:

> … a Transco will be driven, through appropriate incentives[#] to minimize costs, maximize throughput, achieve efficient levels of congestion and reliability, and expand the transmission system when economically justified. Unlike an ISO, this alternative structure will retain the efficiencies gained by integrating the operation of the system with the maintenance, engineering, construction, and restoration of that same system.

The existing ISOs all call for planning processes with central roles for the ISOs themselves. The Texas PUC favors a strong role for the ERCOT ISO in transmission planning because it can do the work "objectively and impartially from a regional perspective." The Texas PUC revised its transmission rule to increase the ISO's role, stating that the ISO "shall supervise ERCOT transmission system planning and exercise comprehensive authority over the planning of bulk transmission projects … ." The PUC is likely to give considerable deference to ISO recommendations concerning the facilities that need to be built, However, the transmission owners retain the obligation to build new transmission that the ISO determines is needed.

The existing ISOs differ in the assignment of transmission costs to generators (both new and existing) and to loads. The ISOs also differ in their use of competition to decide on new projects. The California ISO uses a two-phase process. In the first phase, the ISO develops a statewide transmission plan, based in large part on plans filed by the individual transmission owners. In the second phase, the ISO issues a request for proposals for projects that can meet the state's transmission needs at least cost; these projects can include local generation, demand-side measures, or transmission.

---

[*] In our view, the ISO-vs-Transco debate is not relevant to transmission adequacy because the transmission planning and pricing complexities discussed above apply to *all* regional transmission organizations.

[#] "Appropriate incentives" are the key to make either Transcos or ISOs economically efficient, encouraging them to appropriately balance reliability and commerce.

The interactions among NEPOOL, ISO New England, independent power producers, and FERC show how one region is addressing transmission expansion and also what FERC allows in assigning the costs of new transmission facilities to different entities. The New England transmission system is largely uncongested internally but regularly congested on the interfaces to New York and Canada. The system operators relieve congestion by redispatching generation, with the redispatch costs shared among all loads within the region. New England has a postage-stamp transmission tariff with no locational signals, which provides no information to generators on where to locate new facilities. Historically, generators paid for interconnection facilities, and loads paid for the remainder of the transmission system. As in most regions, generation and transmission were jointly planned through an integrated, regional process dominated by the large investor-owned utilities. Now, with 30,000 MW of new generation being proposed (the existing system has 25,000 MW of generation), new generation is driving transmission expansion.

NEPOOL proposed a process that requires a new generator to be "fully integrated" with load, which means that the new generator must be capable of serving load anywhere in New England without reducing the ability of any existing generator to serve load anywhere in New England. Transmission, therefore, would have to be expanded to accommodate each new generator. The new generator would pay 50% of the cost, and loads would pay the remaining 50%. If the expansion, based on its $/MW of new generation, cost more than the system average embedded cost, the new generator would also pay the full above-average cost.

FERC rejected the proposed process and directed NEPOOL to develop a realistic evaluation process and a system that uses locational prices to manage congestion. FERC's rejection of the NEPOOL proposal suggests there is no way to avoid centralized transmission planning and assessment, based in part on realistic *judgments* concerning future generation-market conditions. FERC viewed as unrealistic and anticompetitive NEPOOL's assumptions that any new generation must be able to reach all loads within NEPOOL, that all existing generation remains online (i.e., new generation displaces no existing generation), and that all prior interconnection applications result in construction of new generation.

FERC's Order 2000 specifies one of the eight minimum functions of an RTO to "… be responsible for planning, and for directing or arranging, necessary transmission expansions, additions, and upgrades that will enable it to provide efficient, reliable and non-discriminatory transmission service and coordinate such efforts with the appropriate state authorities." FERC favors RTOs for transmission planning and expansion because "… a single entity must coordinate these actions to ensure a least cost outcome that maintains or improves existing reliability levels. In the absence of a single entity performing these functions, there is a danger that separate transmission investments will work at cross-purposes and possibly even hurt reliability."

FERC requires the RTO plans to include two key features. First, the process must "encourage market-driven operating and investment actions for preventing and relieving

congestion." FERC strongly favors market mechanisms to deal with congestion rather than the engineering approach embodied in NERC's current transmission loading relief procedures.

Second, the process must "accommodate efforts by state regulatory commissions to create multi-state agreements to review and approve new transmission facilities." FERC believes that the emergence of RTOs might encourage the development of regional regulatory bodies to oversee the certification and siting of new transmission facilities that serve regional, rather than local, needs.

Finally, FERC recognized that a transmission owner that invests in new transmission may be concerned about recovering its investment. Therefore, FERC encourages RTOs to propose pricing incentives that will encourage RTOs to make efficient investments in new transmission facilities. Such incentives could include a higher return on equity, accelerated cost recovery, performance-based rates, or other mechanisms.

SUMMARY

Generation and transmission adequacy pose troublesome transitional and, perhaps, long-term issues. Perhaps the key generation-adequacy problem is the absence of a demand-side response to real-time pricing. Economic theory suggests that consumers and suppliers, in response to real-time prices, will take appropriate steps to ensure generation adequacy. But, if most retail consumers continue to face prices that have little or no temporal variation, this approach will be short-circuited. Until real-time pricing is available to at least some retail customers, traditional approaches to maintaining generation adequacy may be needed.

Different problems arise with transmission, centered about the appropriate institutions that will analyze and plan for transmission enhancements, decide on which transmission alternatives should be built, pay for these investments, and recover the costs of these investments from wholesale and retail customers. In the long run, large regional organizations, as recently proposed by FERC, will likely take on most of these responsibilities, but such organizations now exist only in some parts of the country (covering about one-third of U.S. electricity demand), and these ISOs are very new and still evolving. Finally, the importance of congestion pricing and congestion rights, which could expand greatly the role of competition in enlightening transmission-expansion decisions, is still hotly debated and largely untested. Unlike generation, transmission planning and investment will likely continue to be shared between markets and regulators.

7. CONCLUSIONS

Restructuring the U.S. electricity industry involves the vertical deintegration of the bulk-power system and creation of competitive markets for energy and various reliability services. These changes require operational and institutional changes. Because the electricity industry is so complex, both in terms of its physical characteristics and its organizational structure, making such changes is a multi-year effort. The current transitional state of the industry, part regulated and part competitive, further complicates bulk-power markets, operations, and reliability.

Vertical deintegration will yield three sets of bulk-power entities: competitive generation, FERC-regulated RTOs, and FERC-regulated transmission owners. Transcos combine the second and third entities. Increasingly, private companies will choose to be in the generation business or the transmission business, but not both.

Reliability need not decline in a competitive electricity industry. Indeed, the use of competitive markets to acquire reliability resources should permit the electricity industry to maintain traditional levels of reliability at lower cost.

Although the industry will rely more and more on competitive markets, rather than on central authorities, to deliver electricity to consumers efficiently and reliably, creating and operating such markets is difficult. The real-time nature of electricity production and consumption and the free-flowing nature of transmission lines complicate the use of markets for these functions.

On the other hand, greater reliance on markets affords new opportunities for resources other than central-station power plants. Electricity markets and bulk-power reliability will increasingly be managed by RTOs that have no ties to generation. As a consequence, these RTOs will be largely indifferent to the types of resources (demand vs supply, large vs small) that provide energy and reliability services. In particular, customer response to time-varying electricity prices is likely to play a major role in maintaining reliability and lower electricity costs for all consumers.

APPENDIX: ORDER 2000 ON REGIONAL TRANSMISSION ORGANIZATIONS

On December 20, 1999, FERC issued a major rule on regional transmission organizations (RTOs). Order 2000 is surely the most significant action on bulk-power restructuring since FERC issued its April 1996 Order 888 on open-access, nondiscriminatory transmission service. FERC's new order, although voluntary, strongly encourages jurisdictional utilities to join RTOs by December 15, 2001. The order requires utilities to file reports with FERC by October 15, 2000 with a proposal for an RTO, or, alternatively, a description of the utility's efforts to participate in an RTO and the reasons for not participating in such an entity.

FERC sees RTOs as an important vehicle to address engineering and economic inefficiencies in current bulk-power operations and to reduce opportunities for undue discrimination in provision of transmission services. The Commission believes that RTOs will accommodate competition and improve market performance, reform transmission pricing (including the elimination of pancaked rates), and facilitate lighter-handed FERC regulation.

FERC identified four fundamental characteristics and eight key functions of an RTO. The four characteristics include:

- Independence: the RTO staff and board of directors must be independent of market interests (e.g., generation and retail service), in particular the staff and board can have no financial interests in any market participant, and the RTO must have exclusive and independent authority to propose rates, terms, and conditions for transmission service;

- Scope and regional configuration: the RTO must operate over a region that is large enough to permit it to maintain reliability and support efficient and nondiscriminatory markets for energy and reliability services;

- Operational authority: the RTO must be the security coordinator for its region and it must be able to perform its functions in a nondiscriminatory manner; and

- Short-term reliability: the RTO must have exclusive authority for maintaining short-term reliability (security) for the grid it operates, including exclusive authority over all interchange schedules, redispatch of any generation to maintain reliability, and approval of maintenance outages of transmission facilities.

The eight required functions include:

- Tariff administration and design: the RTO must be the sole provider of transmission services, the sole administrator of its open-access tariff, and have authority to approve requests for new interconnections;

- Congestion management: the RTO must operate markets to manage congestion through the provision of efficient price signals (e.g., locational marginal prices) to all transmission customers;

- Parallel path flow: the RTO must implement procedures to address parallel flow issues within the region, working with adjoining RTOs;

- Ancillary services: the RTO must be the provider of last resort for the six ancillary services specified in Order 888, and the RTO must operate a real-time (intrahour) balancing market;

- OASIS: the RTO must administer an Internet site for its Open Access Same Time Information System for all the transmission facilities it controls, including independent determination of total and available transmission capabilities;

- Market monitoring: the RTO must monitor its markets to identify design flaws, market power abuses, and opportunities for market improvements.

- Planning and expansion: the RTO must be responsible for planning and directing the expansion of its transmission grid, including the use of market-driven operating and investment actions to prevent and relieve congestion, and accommodation of efforts by state regulatory agencies to create multistate agreements to review and approve new transmission facilities.

- Interregional coordination: the RTO must develop mechanisms to coordinate its activities with other RTOs in the region.

To encourage utilities to join RTOs and to ensure appropriate expansion of transmission grids, FERC will consider innovative transmission rate treatment. Such innovative methods might include a rate cap, a risk-adjusted rate of return on transmission investment, shorter depreciation schedules, and performance-based transmission rates. Applications for such innovative rates must show that the rates benefit consumers, not just transmission owners.