Reforming Texas Electricity Markets

If you buy the power, why pay for the power plant?

BY ANDREW N. KLEIT AND ROBERT J. MICHAELS

In the United States and around the world, electricity restructuring is converting regulated monopolies into market regimes. The characteristics of those markets, however, are critical determinants of their performance and remain the subjects of active policy debate. One important issue is whether electricity markets can—without government intervention—provide adequate generation to reliably power society's needs.

Advocates of intervention believe that competitive markets for energy should include additional payments to firms for having reserve generation capacity beyond what ordinary electricity rates would incentivize. Critics of this idea see such "capacity payments" as unnecessary subsidies to electricity producers. Most U.S. regional transmission operators (RTOs), including the Pennsylvania–New Jersey–Maryland Interconnection (PJM) and the New York Independent System Operator (NYISO), operate capacity markets. In California, state regulators impose "resource adequacy" requirements on utilities, but do not directly operate markets. As capacity charges have risen, so has their political salience.

This article reviews the rationales for capacity markets recently proposed for the Electricity Reliability Council of Texas (ERCOT). Some have argued that low levels of investment in generation on ERCOT are reducing reserve margins to levels that threaten reliability. Others believe that ERCOT's "energy-only" regime can suffice to incentivize adequate investment.

Andrew N. Kleit is professor of energy and environmental economics at Pennsylvania State University. Robert J. Michaels is professor of economics at California State University, Fullerton.

Michaels gratefully acknowledges financial support from the Texas Public Policy Foundation. The views expressed in this article are solely those of the authors and not necessarily those of their affiliations.

Why Capacity Markets?

Capacity markets do not exist for goods other than electricity. The dairy industry remains viable without payments by retailers for "cow capacity" on top of milk prices. Any argument for electrical capacity markets should explain why they are needed when such markets are unnecessary or inefficient elsewhere.

The possible rationales rest on two properties of electricity and one economic institution. First, electricity cannot be stored at reasonable cost (except in hydroelectric facilities). Second, a grid operator must match production and demand instantaneously by altering generation or curtailing customers. A surplus of production over demand for a fraction of a second will overload lines, a deficit will produce instability, and either can black out an entire region.

Third, decades of heavy regulation have perpetuated inefficiencies and retarded innovation in power markets. In particular, instead of efficient time-varying prices that reflect the marginal cost of production, most users pay fixed prices that recover average cost. If scarcity blacks out some regions but not others, an economic misallocation will ensue because customers willing to pay a substantial amount for reliable service will lose it, while others whose lights stay on may require small compensation to tolerate some darkness.

Traditional utility regulation made monopoly utilities respon-

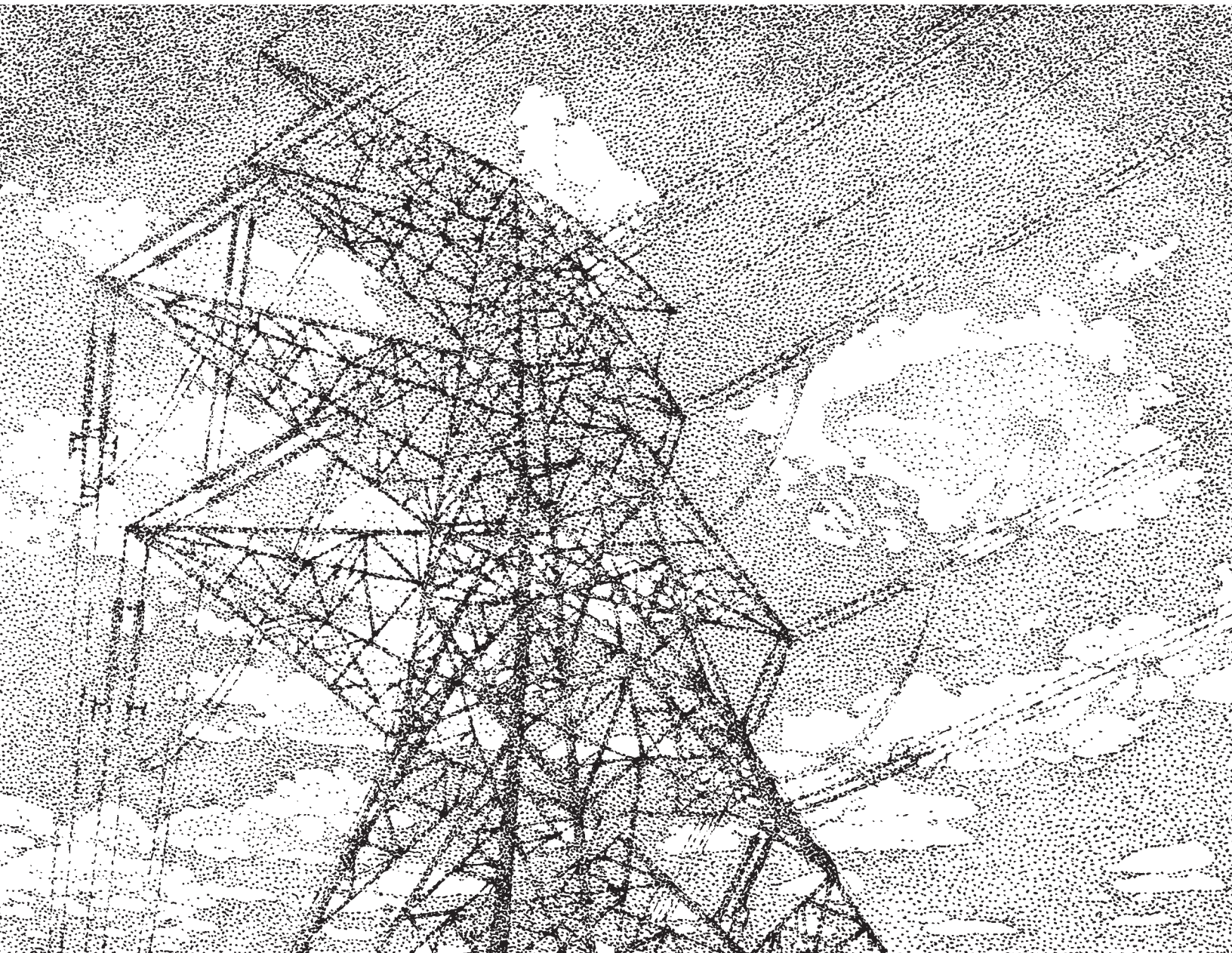
ILLUSTRATION BY MORGAN BALLARD

sible for both day-to-day reliability and prudent investments in generation. Those investments of necessity included generators with high marginal costs that would operate to maintain reliability for only a handful of peak hours. A monopoly utility could fold such costs into its regulated "revenue requirement" without further itemization, secure in the knowledge that regulators would approve higher electricity rates sufficient to cover costs. In an unregulated system, however, it is generally believed that investors in peaking generators can only recover their costs if prices are very high while they are operating.

Aggravating those potential problems, most power markets have regulatory caps on wholesale electricity prices. During system peaks, most available capacity in an RTO market must either produce energy or be committed to providing such "ancillary services" as reserves. With competition thus constrained, a generator may be able to profit by bidding above its marginal cost, particularly if users cannot respond by reducing consumption. The RTO thus faces a dilemma: it wants to foreclose monopolistic profits while ensuring that returns are high enough to induce investment, all in an environment where the number of peak hours cannot reliably be predicted.

Wholesale price caps have been used as a compromise between competitive and monopolistic incentives. PJM's cap, for example, is $1,000 per megawatt-hour (MWh), while ERCOT's is currently $4,500 and will likely rise further under regulations that are now under consideration. The problem with this compromise is that caps can cut generator revenues in times of scarcity and reduce their prospective investment returns. Such shortfalls (in industry parlance, "missing money") constitute another rationale for capacity markets whose payments could make up the difference. Caps as high as ERCOT's, however, reduce the likely disincentives of missing money.

The constellation of generator costs can produce another variant of the missing money argument. In one theoretical model, much of the electricity in a competitive energy market is supplied by baseload generators with lower energy costs and higher capital costs than the peaking units that supply the remainder. If competition sets energy prices at marginal cost, investment in peakers will be insufficient (possibly zero) because they do not earn revenues that allow recovery of capital. The relevance of this



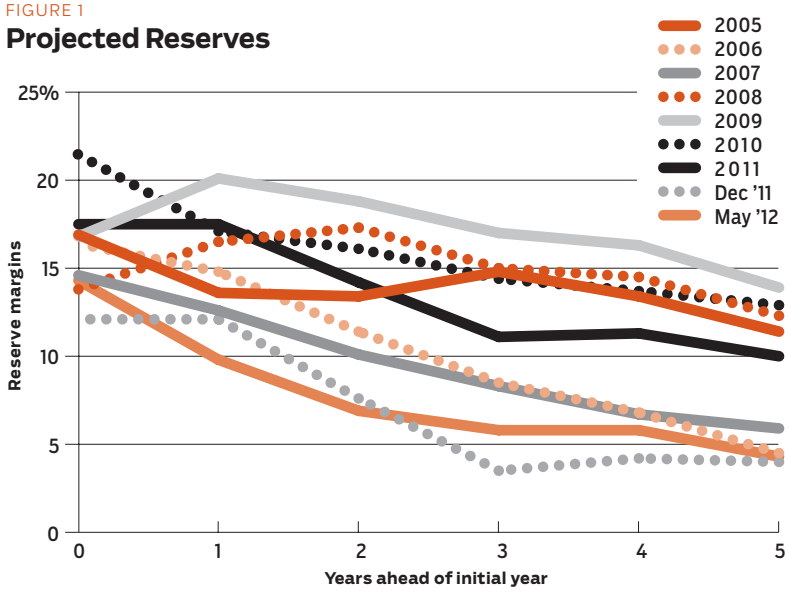
argument, however, depends on its correspondence with reality. The available set of generators in most RTOs appears sufficiently diverse that this version of missing money has been neither an important operational problem nor a deterrent to investment.

Markets generate prices whose movements convey information about shifts in consumers' valuations and producers' opportunity costs. Because prices affect the returns to alternative choices, they induce resource owners to shift toward more profitable activities and consumers to economize on goods whose relative prices have risen. Capacity prices, however, do not emerge from a market. Rather, they come from artificially based "demand curves." These arbitrary curves are not to be confused with textbook demand curves that represent valuations consumers voluntarily place on products. Planners determine these capacity demand curves with the intent of creating a price that will induce amounts of generation investment that the planners deem sufficient. The position and slope of this demand curve are set by RTO committees that need not justify their decisions or relate them to underlying market forces. Its artificiality means that prices in capacity markets will only accidentally be indicators of economic scarcity.

There appears to be general agreement that if the returns to generation investment in Texas are in fact inadequate, that problem lies with a small set of peaking generators and exists for no more than 100 to 200 hours per year. If these units are the problem, any capacity policy should concentrate on them rather than on the much larger set of all generators, many of which are profitable at competitive prices. A capacity market, however, is vastly more complex because it pays all generators (and users who can curtail consumption quickly) rather than just that subset. We show below that there are strong reasons to doubt the common assertion that peaking generators in ERCOT are intrinsically unprofitable without capacity markets.

FIGURE 1

Projected Reserves



Source: Capacity, Demand, and Reserves Report

Markets and Hypotheses in ERCOT

Advocates of capacity markets believe that RTOs without capacity markets display performance inferior to those with them. The most obvious test of this hypothesis would compare retail bills over the long term between the two types of RTOs, but lack of long-term data forecloses that comparison. We can, however, examine two related hypotheses. First, if missing money is relevant, ERCOT should see severe booms and busts in generation investment. There should be relatively long intervals of excess reserves followed by shortages that reflect both investor inertia and construction delays. Second, in years of excess capacity, energy prices in ERCOT should allow recovery of little more than a marginal generator's variable costs. Years of short supply are not predictable and a generator's annual returns will depend on randomness in weather and operating conditions. Because revenue stability can be an important determinant of the cost of capital, extreme randomness of annual returns in ERCOT would be evidence that a capacity market could materially improve performance there. An accurate estimate of those returns, however, requires an accounting for all income sources available to a generator. Unfortunately, as we will discuss below, regulators require that ERCOT's revenue estimates disregard some of those income sources and hence understate actual profitability.

Investment paradox | Some regulators, politicians, and interested generation owners who support a capacity market see ERCOT's recent history as evidence favoring their position. For example, ERCOT's December 2011 Capacity, Demand, and Reserves (CDR) Report predicted summer reserve margins of 12.1 percent in 2012 and 2013, close to ERCOT's longstanding 13.75 percent standard, but saw them falling to only 4 percent by 2014. By May 2012, delays in bringing a new coal-fired generator online and the shutdown of 2,000 MW of coal capacity in expectation of new federal air pollution rules had driven the 2013 estimate below 10 percent. Figure 1 shows that near-term reserve "crises" that fail to materialize have been the rule rather than the exception in ERCOT. Each line shows an annual CDR report's predicted margins for the current summer and for the next five years, based on generators known to be under construction or licensed as of the report date. Perhaps unsurprisingly, the lines almost all slope downward, indicating the reserve situation will worsen significantly as time passes. In reality the 2012 situation was hardly exceptional. Four of the CDR reports since ERCOT's founding (2006, 2007, December 2011, and May 2012) project five-year margins between 4 and 6 percent. In 2011 and 2012, five years after the 2006 and 2007 reports, ERCOT's actual reserves satisfied its adequacy criteria.

These semi-official projections exclude numerous known potential and actual resources, including mothballed fossil fuel capacity, 50 percent of available

external direct current transmission ties, and planned generators under study for interconnection. As of May 2012 these projected 2016 resources totaled 7,409 MW. Even if no other units are built between now and 2016, a relatively small 5,369 megawatts (MW) of additional generation would give ERCOT a 13.75 percent margin in that year, under plausible assumptions about load growth and diffusion of demand management.

The confluence of several unusual events rendered ERCOT's 2012 capacity situation somewhat extreme. Yet, as had also happened before, by the end of summer 2012 the problems had become manageable. A court stay of the new federal air pollution rules brought coal-fired units back into operation, and an additional 2,000 MW of mothballed gas-fired units were brought back into service. A recent report by a member of the Public Utility Commission of Texas (PUCT) found that in 2012 4,318 MW of new generation had been announced, most of which had either obtained financing or started construction. Assuming a lower forecast demand scenario than ERCOT used, these units alone would yield reserve margins of 19.6 percent in 2013, 16.7 percent in 2014, and 13.2 percent in 2018, not counting any additional capacity that might materialize in the interim.

In light of those data, claims of "market failure" by capacity market advocates are hard to take at face value for a number of reasons.

In light of those data, claims of "market failure" by capacity market advocates are hard to take at face value. First, assumptions about future reserves are intended to be conservative and do not adequately account for less-certain power sources that are likely to appear. Second, there is no evidence that investor behavior destabilizes the generation market and that profitable power plants will somehow go unbuilt without capacity requirements. Third, there is no plausible way that a capacity market could have foreclosed the events that led to the shortfalls of early 2012 that resulted from federal regulation and unavoidable randomness. Fourth, policymakers must bear in mind that reliability at high loads depends on adequate peaking, rather than baseload, resources. Finally, as noted above, even if there were valid reasons to subsidize peaking plants, they would not necessarily apply to baseload units.

Ancillary services option | The longer-term growth of both resources and demand in ERCOT appears to indicate that investors in generation are earning adequate returns. By contrast, calculations from ERCOT's market monitor indicate returns that are extremely low. These estimates of "Peaker Net Margin" (PNM) are made pursuant to regulatory formulas that do not include important revenue sources. Officially, PNM equals the difference between the revenue a peaking unit can earn from energy sales in spot markets and its operating cost over a year. (We note that a large majority of actual sales are made via long-term contracts whose terms may or may not adjust to spot market prices.) According to ERCOT, between 2002 and 2007 a new combustion turbine with a gas-to-electric-power conversion factor ("heat rate") of 10.5 would have recovered all of its fixed costs (including interest) by earning a PNM between $65,000 and $80,000 per MW-year. In 2008 and 2009 the estimated lower limit on viability ranged from $70,000 to $95,000 per MW-year and the corresponding range for 2010 and 2011 was between $80,000 and $105,000. A more fuel-efficient (heat rate = 7) combined-cycle gas generator needed to earn at least $100,000 per MW-year in 2003 through 2006 and $105,000 per MW-year in 2007 through 2011. Only rarely have net earnings from energy sales by a peaking plant met or exceeded those amounts.

The PNM calculations suggest that little if any new capacity would appear in ERCOT, but paradoxically capacity has kept up with load. We can resolve this paradox by introducing revenue sources that are excluded from the PUCT's calculation but are relevant for profitability. Most importantly, a generator can choose between selling in the energy ("balancing") market and its ancillary services markets. There are three ancillary services markets:

- ***Regulation reserve*** is generating capacity that follows instantaneous load changes to maintain system frequency of 60 Hertz.
- ***Responsive reserves*** are operating generators available to increase output when a generation or transmission failure occurs in a 10-minute period.
- ***Non-spinning reserves*** ("non-spin") are generators not currently operating ("spinning") that can be ramped to a specified output within 30 minutes, or large loads eligible to act as reserves that are interruptible on 30 minutes' notice.

A generator whose ancillary services bid was accepted allows ERCOT to utilize its capacity as needed. If called upon to operate, the generator receives the balancing market price for its output, which may be above or below its marginal cost. Revenue from ancillary services thus comes in two parts: First, the day-ahead market sets the hourly amount that a successful bidder receives for making its unit available for dispatch. Second, if ERCOT orders that unit to produce energy, its owner receives the real-time balancing market price. If that price is below operating cost, the owner must take the loss.

To simplify the approximation, we assume that the owner can predict with some accuracy today's and tomorrow's balancing energy prices and the probability it will operate. The generator must compare three scenarios:

- It will supply energy if the difference between the balancing (energy) market price and marginal cost is positive and it exceeds the expected net income (weighted by the probability of call) from committing to ancillary services.
- It will supply ancillary services if expected net income from ancillary services is positive and exceeds the net income it would get from energy sold into the balancing market. To be conservative, we restrict ourselves to non-spinning reserves, which typically carry lower prices than other services.
- It will remain idle if both net income from the balancing market and expected net income from ancillary services are negative.

TABLE 1

Generator Operating and Non-Spin Reserve Hours

Year	Heat rate	Probability of being called upon	Energy hours %	Non-spin hours %	Idle hours %
2008	7	0.1	37.26	19.72	43.02
2008	7	0.2	36.66	16.86	46.48
2008	7	0.05	37.52	22.55	39.94
2008	10.5	0.1	8.12	29.16	62.72
2009	7	0.1	28.68	33.82	37.50
2009	7	0.2	27.93	25.55	46.52
2009	7	0.05	29.02	40.83	30.15
2009	10.5	0.1	7.75	30.04	62.21
2010	7	0.1	27.21	30.63	42.16
2010	7	0.2	26.84	26.77	46.39
2010	7	0.05	27.26	33.39	39.35
2010	10.5	0.1	18.05	30.69	51.25

For the years 2008–2010, we perform five PNM calculations that use the market monitor's hourly data but also include revenues from provision of non-spinning reserve services. Four calculations are for a combined-cycle gas generator with a heat rate of 7 (lower figures are better). Its marginal cost per MWh is 7 (i.e., its heat rate) times the market price for a million British thermal units (MMBtu) of gas, plus $4 in variable operation and maintenance costs. We examine the effects of differing probabilities of operation by performing the calculation for probabilities of 0.05, 0.1 and 0.2. The fifth calculation assumes a less efficient combustion turbine whose heat rate is 10.5 and has an operating probability of 0.1.

Table 1 shows the percentages of total annual hours under our various assumptions during which a generator will produce energy, make itself available for non-spin service, or remain idle because neither is profitable. Table 2 shows generator net income for each year under the same assumptions that underlie Table 1. It includes the income viability thresholds for each type of unit in each year as described above. In the exceptional year of 2008, net revenues from energy sales alone are over the viability threshold and adding sales of ancillary services further increases the amount. In 2009, revenues from the energy-only market are in fact insufficient to meet the PNM threshold in all cases, as are those earned when the generator has the additional choice of producing ancillary services.

In 2010, however, generator net income with an ancillary services option in all of our cases falls just short of the profitability threshold. Added up over the three years of data available to us, the unit with a heat rate of 7 is clearly profitable and the unit with a heat rate of 10.5 narrowly achieves viability. These calculations almost surely understate potential net income because they assume that non-spinning reserves are the only possible ancillary service a generator can produce. With more alternatives, a generator could supply other ancillary services that would be more profitable at some times. Unfortunately, data that might refine the calculation were not available to us. It appears clear, however, that generators whose choices more accurately reflect reality will either approach or exceed profitability over time and that investment in ERCOT's energy-only market is, in fact, viable. Thus, the PNM measurement that is designed to model financial revenues for peak generation is revealed to be an administrative creation that does not fully reflect

TABLE 2

Generator Net Income from Energy and Non-Spin Markets

Year	Heat rate	Probability of being called upon	Energy income	Non-spin income	Total income	Non-spin income %	Income viability threshold
2008	7	0.1	$181,905	$17,352	$199,258	8.71	$105,000/ MW-year
2008	7	0.2	180,891	17,491	198,381	8.82	
2008	7	0.05	182,280	17,750	200,030	8.87	
2008	10.5	0.1	121,959	23,279	145,339	16.02	$70,000/ MW-year
2009	7	0.1	$65,751	$12,964	$78,715	16.47	$105,000/ MW-year
2009	7	0.2	65,249	12,302	77,551	15.86	
2009	7	0.05	65,992	13,734	79,727	17.23	
2009	10.5	0.1	47,104	12,990	60,095	21.62	$70,000/ MW-year
2010	7	0.1	$79,482	$20,129	$99,611	20.21	$105,000/ MW-year
2010	7	0.2	78,660	19,497	98,158	19.86	
2010	7	0.05	79,677	20,905	100,583	20.78	
2010	10.5	0.1	57,635	19,852	77,487	25.62	$80,000/ MW-year

the economic opportunities open to generators.

Low PNM estimates in some years also reflect inefficient practices that ERCOT is working to eliminate. Specifically, at peaks when non-spinning reserves must be called, flawed operating procedures until recently brought them into the energy market at an effective bid price of zero. These actions have sometimes produced significant "price reversals" that depress the market price at times when scarcity should be driving it upward. Independent power producer Calpine has assembled a list of 54 incidents of non-spin deployment between December 6, 2010 and May 28, 2011, totaling 184.7 hours. The fall in market price between the last 15-minute period before calling non-spin and the first period after the call averaged $134 per MWh, implying a potential revenue loss to peaking plants of $24,700 over that six-month period. The PUCT addressed this problem (at least in part) in late 2011 by requiring non-spin capacity to be brought into the market at a bid price of at least $120 per MWh.

Demand-side participation | Until recently the price that consumers paid for electricity at any instant might be only distantly related to the marginal cost of producing it, resulting in inefficiencies in the market and threatening reliability. If consumers can see and respond to prices, reliability-based rationales for capacity markets lose their relevance. Most RTOs with capacity markets allow retail suppliers to treat verifiable demand response as a capacity resource for compliance. ERCOT's demand response mechanism allows loads (electricity customers) to participate in four possible reserve services:

- ***Regulation (up and down)*:** Loads that are automatically controllable by ERCOT, which requires telemetry and four-second responses in order to maintain system frequency. Qualifying loads are also eligible to provide non-spin service.
- ***Responsive reserves*:** ERCOT allows up to 1,400 MW of load controlled by telemetry, but they cannot exceed 50 percent of the responsive reserve market. Suppliers must install under-frequency relays with instantaneous response and be able to manually interrupt their loads on 10 minutes' notice.
- ***Non-spinning reserves*:** Loads can participate as non-spinning reserves and must also be callable by telemetry to supply the energy market with small increments of power, known as "droop."
- ***Emergency response services* (ERS):** ERCOT selects qualified loads, generators, and aggregations of loads and generators to supply incremental production and load reductions specifically for deployment in grid emergencies. Auctions take place every four months for supplies that vary from about 200 MW for off-peak hours to about 1,800 MW on-peak. ERS may indeed bring the benefits of more load participation, but it also allows the potentially inefficient payment of different prices to loads and generators.

Pursuant to PUCT policy, the transmission and distribution companies in ERCOT are currently installing "smart" meters for all retail consumers, further increasing the potential for efficient pricing and demand response. We cannot yet project the volume of consumer reaction to the new options, but note that ERCOT is currently preparing for substantial response. If wholesale price volatility increases, this service should become more attractive. Unfortunately, the price distortions inherent in a capacity market or resource adequacy requirement would artificially reduce volatility and blunt incentives for more demand response.

Conclusion

The theoretical case for capacity markets is weak at best. Many of its arguments depend on oversimplified assumptions that are at variance with reality, particularly those that are necessary to produce the missing money phenomenon. The lack of demand response is becoming less relevant as markets develop, more users see prices based on marginal cost, and demand management becomes more widespread. Prices that prevail for capacity and amounts to be invested in it will be administratively set and have only a tenuous connection with economic efficiency.

An examination of ERCOT's current state does not provide coherent support for capacity market advocates. The 2011–2012 capacity shortfalls cited by ERCOT critics are largely idiosyncratic—the results of unusual political, regulatory, and weather events. In many years, a highly conservative three- or five-year projection would show ERCOT falling dangerously short of reserves, but market forces have invariably succeeded in restoring their generation adequacy. Claims by critics that investment is persistently unprofitable in ERCOT's energy-only markets rest on a regulator-determined formula (Peaker Net Margin) whose definition is restricted to only a subset of all potential revenues.

Two remaining barriers to efficiency and reliability are in the process of falling. The first is the set of rules that lower market prices during peak periods when they should be raised, which the PUCT continues to address. The second is demand management that has yet to grow the institutions and attain the scale that would make it truly symmetric with supply in setting prices. These problems, however, do not stem from any inherent flaws that render energy markets incapable of functioning efficiently and properly without capacity markets. R

READINGS

- "Economics and Design of Capacity Markets for the Power Sector," by Peter Cramton and Axel Ockenfels. University of Maryland working paper, Oct. 30, 2011.
- "ERCOT Investment Incentives and Resource Adequacy," prepared by the Brattle Group. June 1, 2012.
- "Money for Nothing in the Power Supply Business," produced by the American Public Power Association. *Issue Brief,* March 2012.
- "Price Responsive Load: Next Steps—Data Collection," by Paul Wattles and Karen Farley. Electricity Reliability Council of Texas, October 16, 2012.
- "Ready or Not, Here Comes the Smart Grid," by Seth Blumsack and Alisha Fernandez. *Energy*, Vol. 37, No. 1 (2012).
- "Resource Adequacy in ERCOT," presentation graphics submitted in PUCT Project No. 40000, by Kenneth W. Anderson Jr. Nov. 12, 2012.
- "Texas Nodal Market Guide, Version 3.0," published by the Electricity Reliability Council of Texas. December 2010.