Electricity Restructuring

The Texas Story

L. Lynne Kiesling and
Andrew N. Kleit, Editors

The AEI Press
Publisher for the American Enterprise Institute
WASHINGTON, D.C.

4

Achieving Resource Adequacy in Texas via an Energy-Only Electricity Market

Eric S. Schubert, Shmuel S. Oren, and Parviz Adib

Assuring a reliable supply of electricity in a market-based system has been a central concern in restructured electricity markets throughout the world and the subject of an ongoing debate among academics, industry leaders, and policymakers. The main problem is how to reconcile engineering criteria for reliability and resource adequacy with market mechanisms that will provide price signals for investment, while satisfying regulatory concerns regarding just and reasonable costs for consumers.

Three approaches to ensuring generation adequacy currently exist. The first uses energy-only markets with limited price mitigation (for example, high caps on offers into market) that rely on energy remuneration and scarcity pricing to guide investment. The second uses adequacy mechanisms based on capacity products, which take two forms. One is capacity payments to installed or operational capacity, such as are used in Spain, Italy, Korea, and several Latin American countries. The other is capacity obligations imposed on load-serving entities (LSEs) that can be met in several ways, including bilateral contracting with regulatory verification, as in California; centralized capacity markets, as ISOs in the northeastern United

The aauthors would like to thank the editors of this book for thoughtful suggestions on early versions of this chapter and Felicia Schubert for helping us prepare the manuscript for publication. The opinions expressed in this paper are not necessarily those of APX Inc. or BP Energy Company. The authors were on the team at the Public Utility Commission of Texas (PUCT) that developed the energy-only resource adequacy mechanism in Texas. Adib and Schubert were on staff at the PUCT at the time, while Oren served as senior market advisor to the PUCT.

States; and combinations of bilateral contracting with bulletin-board trading of standardized contracts or a central capacity market. The third approach, which can be viewed as a market-friendly version of traditional integrated resource planning, is based on central resource procurement that can take the form of competitive tendering through either a request for offers (RFO) process or bilateral negotiation, as in France and some other European countries; or strategic reserve contracts between ISOs and critical resources, as in the Nordpool countries.[1]

In this chapter, we examine the question of which of these approaches is the most effective through the prism of the energy-only resource adequacy mechanism of the Electric Reliability Council of Texas (ERCOT). We begin with a review of the intellectual and policy debate concerning resource adequacy, followed by an overview of the political economy background and the evolution of the energy-only approach to resource adequacy in ERCOT.[2] Next is a discussion of how the ERCOT market design has met the conditions that make an energy-only market workable in terms of controlling market power abuse and enabling suppliers to collect legitimate scarcity rents. We also describe some ongoing efforts to reduce the tension between, on the one hand, engineering procedures focused on reliability objectives that tend to mute scarcity price signals and, on the other, the market goals of providing scarcity price signals that are needed to encourage investment. Finally, while we defend the decision made by Texas regulators to take the path of an energy-only electricity market, we acknowledge that a few more years of operation will provide the empirical basis for further evaluation of the effectiveness of Texas's approach to addressing the complex issue of resource adequacy.

The Resource Adequacy Debate

For over a decade, academics, industry leaders, and policymakers have debated whether capacity mechanisms separate from energy markets are needed in restructured electricity markets, whether such capacity markets need to be centralized, and, if they are centralized, how they should be designed. Some argue that, given their technical, political, and social realities, electricity markets need to be supplemented by some capacity mechanism

that will ensure generation adequacy.[3] The primary objective is to create sufficient incentives for efficient investment choices. In most cases, however, this goal is interpreted as inducing investment in generation that will meet prescribed reliability criteria based on technical rather than economic considerations. Capacity mechanisms, according to this view, would stabilize generators' income streams suppressed by offer caps that are too low to allow generators to recover their fixed costs, restoring the so-called missing money.[4] Capacity mechanisms are also often viewed as a means of achieving efficient investment. From an engineering perspective, capacity remuneration is a mechanism of choice, since it is a "top-down" approach that supports the setting of capacity targets through centralized integrated resource planning and remuneration of the resources on a cost-accounting basis rather than on a market-value basis.

The need for capacity remuneration in addition to payments for energy and reserve provision in the power industry is often rationalized on the grounds that electricity is a necessity and, hence, commodity prices must be controlled, and supply adequacy ensured, through regulatory intervention.[5] Proponents of the capacity approach also argue that *reliability* of supply, which has public-good characteristics similar to national security or fire protection, is a product distinct from *energy,* and, thus, it needs to be regulated and paid for through capacity remuneration.[6]

On the other hand, from an economic point of view, the notion of capacity payments as a mechanism for cost recovery and supply adequacy assurance is an anomaly, unique to the electric power industry, and it originates in the legacy of that industry as a regulated monopoly. In any other industry, however capital-intensive it may be, suppliers assume investment risk and have the opportunity to recover their costs and make profits by selling the commodity or service at competitive market-based prices, while the customers for the service (for example, LSEs) assume price risk. The two sides manage their mutual risk through long-term bilateral contracting between them. This "bottom-up" approach, which treats electricity as a commodity and creates markets without a centralized planning mechanism, has been adopted in Australia, the Canadian province of Alberta, New Zealand, and ERCOT.

From a theoretical economic perspective, the most important question is whether a competitive energy market without separate capacity remuneration

can produce a socially efficient technology mix and total capacity level.[7] The answer to that question is yes, provided that the market is truly competitive so that generators behave as price-takers, and energy prices are allowed to reflect scarcity rents when supply is short. Indeed, it can be shown that when the electricity market is at its optimum in terms of technology mix and total capacity, and the real-time electricity grid is optimally dispatched (transmission constraints notwithstanding), then paying a single clearing price for all the energy produced at each point in the real-time market at the marginal cost of the most expensive unit dispatched will result in a revenue shortfall for all dispatched units. The shortfall is exactly equal to the capacity (fixed) cost of the peaking unit. The framework above assumes some explicit or implicit auction with a single clearing price for all dispatched units where the owners of those units reveal their true marginal costs. It also can enable price-sensitive demand response at the retail level.

The shortfall resulting from marginal-cost pricing based only on generation cost can be recouped, however, without the need for capacity payments by allowing scarcity prices to be set by demand response (at the value of lost load, or VOLL) whenever generation capacity is exhausted.[8] When timely demand response is not technically feasible, it can be approximated by administratively setting the uniform clearing price to an estimated VOLL whenever demand is curtailed due to insufficient supply offers in the cost-based, uniform-price auction. Under such a scheme, the amortized cost of a one-megawatt combustion turbine [CT] per hour equals VOLL per MWh times the loss-of-load probability (LOLP), which is the condition for socially optimal capacity in the system.

Although such a scheme can be implemented even in the absence of active retail demand response, it would obviously benefit from active demand participation, which would provide a market-based VOLL instead of an administrative estimate.[9] The central challenges in implementing such a scheme are, therefore, ensuring a workable level of competition in the market so that generators are not in a position to exert market power on a sustained basis and, second, ensuring that prices will reflect scarcity conditions. Satisfying these two conditions simultaneously is not easy, since market-mitigation schemes used to ensure competitive prices often tend to suppress scarcity prices. Furthermore, shortage conditions are often masked, and the scarcity rents muted by, the system operator's deployment

of reserves and by out-of-market actions aimed at maintaining system reliability and avoiding involuntary load curtailments.

The Political Evolution of the ERCOT Market

The resource adequacy question in Texas arises in a policy context that has brought about more competition within the state's electricity market. The major factors in this context are certain landmark decisions made by the state legislature and the corresponding implementation actions taken by the Public Utility Commission of Texas (PUCT) through its policies and substantive rules.

In the 1990s, the Texas legislature passed two major electricity restructuring bills. Senate Bill 373 (SB 373), passed by the seventy-third session of the legislature in 1995, opened the state's wholesale electricity market to competition with the understanding that any existing wholesale contracts would remain intact until the end of their terms and conditions.[10] Senate Bill 7 (SB 7), passed by the seventy-sixth legislative session in 1999, amended the Public Utility Regulatory Act (PURA) to allow retail competition to begin on January 1, 2002, in areas served by investor-owned utilities within the power region of ERCOT. Municipal and co-op utilities could choose to opt into competition.[11]

The PUCT also took several actions with significant consequences during this period. In general, these actions fell into three categories:

- Actions taken after the passage of SB 373 in 1995 resulted in the establishment of rules necessary to create a level playing field for all participants in the wholesale electricity market. These rules provided nondiscriminatory access to the transmission system and defined terms and conditions for interconnection to the transmission grid by new power sources.[12] In brief, easy interconnection of generation encouraged aggressive investment in new transmission and allowed socialized payment by all loads for new transmission. These were the main factors contributing to significant new generation additions in the ERCOT power region or wholesale market.

- Actions taken after the passage of SB 7 in 1999 resulted in the establishment of rules necessary to create a level playing field for all participants in the retail electricity market. In addition, the PUCT finalized market rules to set parameters for the operation of the wholesale electricity market within the ERCOT power region as a single-control-area operation. These included rules to unbundle integrated electric utilities functionally, to address the treatment of stranded investment, to define a code of conduct for regulated utilities, to define customer rights and protections, and to set design parameters and protocols for the wholesale market.
- The PUCT took a firm stand on discouraging the construction of any new power plants by regulated utilities that could ultimately result in rate-base treatment of such additional investments. Such plants might increase the magnitude of stranded investment and discourage competitive suppliers from entering the incumbent utilities' service territories, which would ultimately harm prospects for developing competitive markets and customer choice.

As a result of these actions, competitive electricity markets have been growing within ERCOT, and independent power producers (IPPs) have gained significantly in their share of installed capacity. The ERCOT power region has had more than ten years of wholesale competition, accompanied by more than seven years of retail competition. Retail choice is available to all customers within the service territories of traditional investor-owned utilities, with no regulatory price protection as of January 1, 2007, and no apparent need for such protection. The Texas retail market is routinely ranked among the top competitive retail electricity markets in the world and the best in North America based on a number of factors, including switching rates by retail customers among competing providers.[13]

The increase in the IPPs' share of installed generation capacities within Texas has been a blessing. Between 1995 and April 2009, approximately 43,000 MW of new capacity was added in Texas, of which IPPs and other nonutility entities, such as combined heat and power producers (CHP), accounted for more than 85 percent.[14] Electric cooperatives and municipal

utilities, which chose not to be competitive at the retail level, and investor-owned utilities accounted for the rest. In terms of power regions, about 90 percent of the new capacity was built in ERCOT, and the rest was divided between the Southwest Power Pool (SPP) and Southeast Electric Reliability Council (SERC).

How the PUCT Chose the Energy-Only Approach

In 2001, the PUCT began a rulemaking on resource adequacy. As part of the proceeding, the commission's staff and the stakeholders reviewed existing resource adequacy mechanisms, including the installed-capacity (ICAP) markets. Following the lead of other established electricity markets in the United States, the staff proposed a centralized capacity market in 2002. Generation owners strongly favored capacity mechanisms, while retailers and industrial consumers were opposed.

In 2003, with the debate over zonal versus nodal market design dominating the stakeholder process and a very large reserve margin in the energy market, the PUCT decided it could postpone consideration of a resource adequacy mechanism. The commission staff and a number of stakeholders noted that postponing the decision on resource adequacy would allow the PUCT to review the outcome of the Federal Energy Regulatory Commission's (FERC's) standard market design process, which might provide a capacity mechanism that the PUCT could adopt.

Despite having recommended the use of a capacity approach to resource adequacy for ERCOT, the staff had a number of concerns about a resource adequacy mechanism based on capacity payments. First, peaking and baseload units had two very different payment streams, which were not easily reconciled in a capacity mechanism. Second, the locational and operating characteristics of the wholesale electricity market were not easily covered in a capacity mechanism. Third, retailers and industrial customers fiercely opposed capacity markets, believing that capacity payments had not proved effective in adding new generation to other markets. And, fourth, unlike electricity markets on the East Coast, ERCOT by 2003 had already functioned as an energy-only market and had attracted substantial new investment without capacity payments.

During the suspension of the rulemaking on resource adequacy, the commission staff had intended to look to the existing markets in the Eastern Interconnection—as it had with the nodal market design—for a solution to the resource adequacy issue. The flaws in the existing ICAP markets had become increasingly apparent, however, and no proven model of capacity markets was available.[15] Even worse, from the staff's point of view, a review of a draft of the Pennsylvania–New Jersey–Maryland (PJM) reliability pricing model (RPM) showed that, far from developing a market-friendly solution to resource adequacy, PJM was moving in the direction of a centralized integrated resource planning approach.[16]

At about the same time, the PUCT commissioners were still reviewing alternatives to a nodal market design, holding two workshops on the subject in December 2004.[17] Afterward, with doubts still lingering about the nodal design, the staff reviewed two existing zonal designs: the United Kingdom and Australian electricity markets. The United Kingdom design was quickly dismissed as an alternative, as it allowed the grid operator to contract actively for resources to counter the position of market participants when those positions were deemed unfavorable to the market. Such an approach was completely contrary to the one ERCOT stakeholders had taken in designing the ERCOT market, which relied on the grid operator's maintaining system reliability without considering the impact of its actions on clearing prices.

Growth in load and the retirement and mothballing of a number of inefficient gas-fired plants caused the commission staff to restart the resource adequacy rulemaking. In February 2005, the staff held a lengthy conference call with the monitors of the Australian market about the various elements of the Australian market design. While it became evident that the Australian zonal approach would not easily be transferred to the ERCOT market, the staff found in the Australian New Electricity Market (NEMMCO) an energy-only resource adequacy mechanism that had been working successfully for more than six years.[18]

The evidence of a working energy-only market provided the stimulus to the commission staff and the needed reassurance to the commissioners to pursue such a solution. The staff began drafting a white paper later that month to explore practical alternative approaches to addressing resource adequacy effectively. Because the Wholesale Market Oversight (WMO) group

(formerly known as the Market Oversight Division) of the PUCT was hosting the semiannual Energy Intermarket Surveillance Group in April 2005,[19] the staff decided to conduct a workshop after the meeting so market monitors from the United States, Canada, and Australia could make presentations on the two competing approaches to resource adequacy: an energy-only market versus separate energy and capacity markets.

In the week prior to the workshop, the staff filed a white paper with the commission,[20] explaining how an energy-only approach might work in ERCOT. It also took the position that an energy-only wholesale market design combined with an active retail market would accelerate adoption of potential innovations, such as market-based demand-side response, as well as upcoming technologies, such as advanced metering and solar power. Expressed in the paper were concerns regarding the need to develop market-based demand response at the retail level, the determination of the offer-cap level, and the question of distinguishing between scarcity pricing and the exercise of market power.

The workshop in April 2005 was the watershed event in the movement away from a capacity approach to resource adequacy that was being considered in other U.S. electricity markets at that time. The Australian regulator made a persuasive presentation, showing that an energy-only market with active retail competition was not only feasible, but had been thriving for more than six years. The Alberta regulator showed an energy-only approach that was being successfully used in North America.

In a subsequent workshop, the commissioners allowed proponents of energy-only markets and capacity markets (who were ERCOT stakeholders) to make a case for each approach. At least one commissioner worried that capacity payments were subsidies that, once established, would be very hard to remove.[21] The capacity market proponents had difficulty coalescing around a single, workable structure, foreshadowing the great difficulties and controversies that would be faced in gaining a broad and stable consensus on a workable capacity mechanism, as was seen in the still controversial development and operation of PJM's RPM and the New England ISO's Forward Capacity Market (FCM). Not surprisingly, owners of large-generation fleets favored a capacity resource-adequacy mechanism, and retailers preferred an energy-only approach. Owners of smaller generation fleets were split on the issue but were concerned that the offer cap

developed as part of an energy-only resource-adequacy mechanism would not be high enough to be sustainable.

By the spring and summer of 2005, the commission was faced with a fundamental choice with respect to the evolution of the wholesale market design and the continued success of retail competition: The PUCT could increase reliance on markets (through an energy-only approach), or it could return to the days of integrated resource planning (through the capacity approach).

The energy-only approach would create the potential for higher and more volatile prices during times of scarcity. Retail load aggregators, such as load-serving entities, would need to learn the skills to manage this price risk effectively. Generators, on the other hand, would face more investment risk than any other market in the United States. The symmetrical price and investment risk, if placed into the design properly, would provide strong incentives for innovation and longer-term bilateral contracting.

Within a few months of the April 2005 workshop, the commission chose the energy-only approach and put its staff in charge of drafting the details. Given the strong similarities between the Australian and ERCOT markets and the proven success of the Australian approach, the staff based its resource adequacy rule on the Australian resource-adequacy mechanism, with some deviations reflecting differences between the two markets.[22]

The choices of the energy-only elements for the ERCOT electricity market were also based on the prevailing political realities within ERCOT. ERCOT had a unified regulatory regime in place, where the PUCT was responsible for regulation of the wholesale and retail markets, as well as the transmission grid. In addition, retail competition had been highly successful in ERCOT, without any requirements that competitive retailers demonstrate to either ERCOT or the PUCT that they could meet their projected needs far in advance of the real-time market.[23] The ERCOT market had been relying on price risk for choices in the scheduling and contracting of resources.

Market power abuse, particularly through physical withholding of resources, also was a serious reliability concern because ERCOT, like the Australian market, was a medium-sized, isolated interconnection with a single settlement market,[24] where sizeable system-wide frequency deviations due to loss of power resources were not uncommon. As a result, the energy-only market needed to provide incentives for owners of generation to offer their resources freely into the real-time market to obtain scarcity pricing,

rather than withhold resources to obtain scarcity pricing through an administrative pricing mechanism.

The remaining challenge the PUCT faced in developing the recently adopted energy-only resource adequacy mechanism was the incorporation of a complementary market power mitigation regime.[25] Market power and resource adequacy intersect on the vexing issue of scarcity pricing. Failure to address market power results in prices that are too generous for producers based on their exertion of market power, producing price signals that do not truly reflect demand and supply conditions. In the long run, the artificially high prices resulting from the exertion of market power are unsustainable; they undermine economic efficiency, weaken public confidence in the market, and create an uncertain regulatory climate that hinders investment in new generation. Similarly, too much price mitigation results in prices that are too generous for consumers, blocking a price signal reflecting actual demand and supply conditions. Low prices discourage generators from developing new generation to meet growing demand for electricity and from replacing the older, less efficient generation available to run the handful of hours needed each year to meet annual peak demand.

Both of these outcomes result in a lack of adequate investment in merchant plant development, and they can cause shortages in power supply. It is, therefore, important for policymakers to address both issues—market power and resource adequacy—at the same time, to ensure that their interdependencies are adequately addressed. Accordingly, the PUCT combined the ongoing resource-adequacy and market-power rulemakings into a single proceeding.[26]

During the deliberations on the resource-adequacy rulemaking, all three commissioners stated repeatedly that increases in the offer cap had to be accompanied by more rapid disclosure of information affecting the operation of the ERCOT markets. The proponents of increased disclosure argued that it would help ensure heightened market transparency and discourage market-power abuse. Increased transparency would also reassure the public that the price changes they observed were the result of a properly functioning competitive market.

The interrelationship of higher offer caps and reduced mitigation, on the one hand, and the more rapid disclosure of information about resource-specific offers, on the other, was consistent with disclosure policies in other U.S. and

foreign markets. In 2006, when the resource-adequacy rule for the ERCOT market was being considered by the PUCT, FERC-jurisdictional markets such as PJM and the New England ISO released resource-specific information six months after they gathered it, which might have been adequate when individual resource offers were heavily mitigated through conduct and impact tests, and offer caps were relatively low. More rapid disclosure of resource-specific information appeared to provide limited benefit under these circumstances in these markets because prices were mitigated in advance of being announced, and those rare circumstances when price spikes occurred were almost always known to market participants.

In contrast, an energy-only resource adequacy mechanism with lighter mitigation of resource-specific offers was less predictable and less transparent. It was therefore argued that more rapid disclosure of resource-specific offers was needed to provide market participants with the same range of information and protection found in FERC-jurisdictional markets. This combination of lighter mitigation and quicker disclosure was already seen in established electricity markets outside of the United States: The Australian electricity market disclosed resource-specific offers with the names of the generators making the offers within twenty-four hours; the New Zealand electricity market disclosed the same information within fourteen days; and the Alberta electricity market displayed the output of each generator, by name, on its website in real-time. Yet it was still debatable whether disclosure of aggregate offer curves would be sufficient to support competition, and whether early disclosure of more detailed information would serve primarily as a means of market mitigation, de facto creating a "shame cap" that might deter the exercise of transitory market power by exposing to scrutiny those firms engaging in it.

Details of the ERCOT Energy-Only Market

The ERCOT energy-only market has a number of features that are unique to U.S. wholesale electricity markets. Among them are offer caps above $1,000 and quick disclosure of offers into the day-ahead ancillary services and real-time energy markets of ERCOT, as well as system-wide market power mitigation divided into two parts: an exemption from system-wide

market power mitigation for generation owners whose fleets comprise less than 5 percent of installed capacity, and a voluntary mitigation plan, to be agreed upon by the PUCT staff and the independent market monitor and approved by the PUCT commissioners, that could serve as a safe haven for the owners of the largest generation fleets within ERCOT.

Higher Offer Caps and Scarcity Pricing. One of the PUCT's broad policy objectives in adopting an energy-only resource adequacy mechanism was to provide greater assurance that generation companies and developers would invest in the resources needed to supply the electricity needs of ERCOT customers by allowing prices to rise in response to scarcity of resources in the market. It would, in particular, encourage the development of such alternatives by providing incentives for the development of new peaking capacity.

The PUCT reasoned that a $1,000 per MWh offer cap would provide sufficient incentive for market participants to build and to contract for new baseload, intermediate, and intermittent renewable generation—resources that could meet about 98 percent of the electricity needs of ERCOT. ERCOT stakeholders, PUCT staff, and the commissioners did not believe, however, that a $1,000 offer cap would necessarily provide incentive for the market to support the 2 percent of hours when electricity demand was at its highest: late weekday afternoons in the summer. New peaking generation or demand-side resources might need an opportunity to earn more than $1,000 per MWh (that is, to earn scarcity pricing) in the ERCOT market to cover their capital costs during this very limited number of hours.

Allowing scarcity pricing would provide load-serving entities with the incentive to procure sufficient peaking generation or demand resources as a hedge against scarcity pricing in the ERCOT spot market during this time. Concurrently, ERCOT would need to establish the appropriate credit limits on load-serving entities to limit their ability to lean heavily and consistently on ERCOT spot markets to meet their customers' demand for electricity. Such a strategy would be risky in the face of scarcity pricing, potentially causing a default in payments to ERCOT that other load-serving entities would need to cover. Prudent credit policies also would provide strong incentives for load-serving entities to bring sufficient generation and demand resources to the ERCOT spot market during annual peak demand

to maintain reliable operation of the ERCOT grid without the need for out-of-market actions by the ERCOT grid operator or a capacity resource-adequacy mechanism.

Another reason the PUCT chose an offer cap higher than the prevailing $1,000 per MWh was its belief that, under an energy-only resource adequacy mechanism, ERCOT could not rely on a daily "must-offer" requirement or capacity payments to ensure the availability of sufficient resources in those situations. A higher offer cap could provide strong incentive for investment in quick-start generation and load response to meet demand in unusual market situations. Such an incentive turned out to be critical in maintaining reliability in ERCOT, which is a small electrical interconnection compared to the Eastern or Western Interconnections in the United States.

As outlined in the approved rule, on March 1, 2007, the offer cap rose from $1,000 per MWh to $1,500 per MWh.[27] On March 1, 2008, it rose to $2,250 per MWh, and it is scheduled to rise again to $3,000 per MWh two months after the market begins operation under a nodal market design, projected to occur sometime in early 2011. The PUCT chose a significantly lower offer cap than its counterpart in Australia, in part because the ratio of all-time peak to average summer-peak demand in ERCOT was not as high.[28] The PUCT decided to phase in the increase over a three-year period, rather than implementing it immediately, consistent with the three-year time frame in the rulemaking to improve market transparency gradually.

The PUCT also decided that to make the offer caps sustainable, ERCOT needed to increase the price-responsiveness of load in spot markets. The commission stated in other PUCT rulemaking projects that the price elasticity of demand was limited by the lack of interval metering for many loads and plans. The PUCT also considered requiring ERCOT to enable advanced meters for residential and other small loads to provide customers and retailers with more discrete electricity usage information than is provided by monthly billings and average-load profiles.[29] The PUCT has completed Project No. 34610, *Implementation Project Relating to Advanced Metering*, for which the commission staff worked with ERCOT stakeholders to develop a plan to implement the back-office infrastructure, as well as settlement software to allow for fifteen-minute settlement of all competitive loads using advanced meters. At the end of 2008, the PUCT also approved plans proposed by the two largest transmission companies in ERCOT, Oncor and

Centerpoint, to deploy advanced meters to all residential and small commercial customers in their territories within roughly five or six years.[30]

Publication of Resource-Specific Offers into ERCOT-Procured Markets. Effective March 1, 2007, most of the required disaggregated information in the ERCOT market was to be disclosed ninety days after the day for which the information was accumulated—that is, within one-half of the previous disclosure time frame of 180 days. This rule was designed to shorten the disclosure period to sixty days, then to thirty days, on the dates when the offer cap was to be raised from the original $1,000 per MWh to $2,250 per MWh to $3,000 per MWh. The implementation schedule for disclosure was tied to the schedule for increases to the offer cap because, throughout the debate in the rulemaking, all three commissioners emphasized the interrelationships of these two issues. They believed the potential for higher prices would require greater assurances to the public that the prices were the result of a competitive market and not of market manipulation.[31]

One major exception to this disclosure schedule pertained to offer curves by individual resources for balancing energy and ancillary services. Because these two areas raised the greatest concerns about the possibility of market power abuse and other market manipulation, the PUCT stated that expedited disclosure of offer curves for these services was appropriate to provide greater transparency to the public and affected market participants. The initial disclosure provisions in the PUCT rules were contested in court by some participants, eventually leading to a compromise that balanced some of their concerns about disclosure of strategic business information against the greater need for public scrutiny.[32] In its final ruling on the matter, the PUCT decided that, as a general rule, the offer curves should be disclosed sixty days after the day for which the information was accumulated.[33]

As part of the disclosure rules, market price-setters would now be identified after forty-eight hours for each settlement interval. For each period that it ran a balancing energy auction or an ancillary-capacity service auction, ERCOT would publicly identify the supplier with the highest-priced offer accepted, along with the price of the offer. This disclosure would be unremarkable and uninformative most of the time. When prices ran high, however, the public would quickly know whose offer caused the price to clear where it did. A supplier would still be able to price its offer however it wanted (up to the prevailing offer cap), but an offer priced significantly above marginal cost would draw public attention if it ended up setting the market-clearing price. Through the threat of intense public scrutiny of any inappropriate market behavior, this targeted transparency was intended to deter persistent gaming without compromising a supplier's ability to offer energy or capacity at prices sufficient to cover a unit's marginal cost.

Moreover, to enhance market transparency further when significantly high prices were offered, the commission approved an event trigger to be used to identify portions of participants' offer curves that should be disclosed after seven days. The event trigger was defined as a calculated value for each interval that was equal to fifty times the Houston Ship Channel natural gas price index for each operating day, expressed in dollars per megawatt-hour (MWh) or dollars per megawatt per hour (MW/h).[34]

Scarcity Pricing Mechanism (SPM). The SPM, based on the Australian model, was intended to raise offer caps to encourage resource adequacy while preventing excessive transfers of wealth from load to generation during years when reserve margins were thin. The rule relied primarily on high energy offers by generators or by demand resources to set the scarcity prices during shortage periods. While such scarcity prices were justified and necessary for cost recovery in an energy-only framework, time lags in construction of new capacity and limited demand-response capability might result in prolonged periods of high prices and "excessive" recovery. Allowing such excessive recovery would result in an unwarranted transfer of wealth (at least from the PUCT's point of view) to generators from load.

The SPM would operate on an annual resource adequacy cycle, in which the peaker net margin (PNM) would be calculated as the sum of all positive differences between the clearing price in the ERCOT real-time energy market and the estimated marginal cost of operating a generic peaker with a heat rate per MWh of 10 million British thermal units (MMBTU). At the beginning of the annual resource adequacy cycle, the system-wide offer cap would be set at one of the offer caps listed above, which was denoted as "high cap" (HCAP). If the PNM were to exceed $175,000 per MW during an annual resource adequacy cycle, the system-wide offer cap would be reset at a lower level, denoted as "low cap" (LCAP),[35] for the remainder of that cycle. The

offer cap would be restored to the highest level allowed by the rule at the beginning of the next cycle.

Exemption on System-Wide Market Power Based on Installed Generation Capacity, or "Small Fish Swim Free." As explained above, the success of an energy-only market hinges on competitive offers that are not inflated through sustained market power abuse, and on scarcity pricing that reflects shortage conditions. This desired outcome traditionally has presented U.S. economists and policymakers with a dilemma. Economic theory suggests that price-taking behavior results in short-run marginal-cost pricing in the real-time market. Short-run marginal costs will not, however, allow for sufficient inframarginal profits to support peaking gas-fired generation that needs to recover a large amount of its fixed costs in the small number of hours it operates in a given year. Unfortunately, those hours when inframarginal profits will be needed are also the times when a number of generation suppliers can exert market power and potentially inflate market prices. To restrict artificially high prices resulting from the exertion of market power (both local and system-wide), U.S. electricity markets have included market mitigation on offers from generators. When offers from *peaking* generation are restricted by *ex ante* mitigation to reflect the short-run competitive outcomes, as has been done in FERC-jurisdictional wholesale markets, the result is inconsistent with scarcity pricing and inframarginal profits. Thus, when exploring the possibility of using an energy-only resource-adequacy mechanism, economists and policymakers confront the "Gordian knot" of having to determine conditions under which scarcity pricing is a function of the exercise of market power or of genuine resource scarcity.

The "bottom-up" alternative that had been used in the Australian market—light or no *ex ante* mitigation of energy offers from generation—has relied on transitory (but not systematic) market power of pivotal suppliers and hockey-stick bidding strategies during shortage conditions, to set scarcity rents through high offer prices. Such an approach had been working in Australia because generation ownership has been sufficiently dispersed among market participants to make the exercise of market power transitory under limited conditions that correspond very closely to conditions of genuine scarcity.[36] Depletion of operating reserves is another common metric for

scarcity conditions used by many U.S. independent system operators, but it would not be useful in Australia, which does not have separate markets for operating reserves.

The Australian approach was contrary to the current stance of policymakers in FERC jurisdictional markets and the theoretical frameworks of leading U.S. electricity economists (almost all of whom have favored a "top-down" approach to resource adequacy).[37] As a result, it was subject to controversy or outright dismissal as a viable alternative for needed scarcity pricing in other U.S. electricity markets. The dismissal of the Australian approach may have reflected the fact that a number of preconditions underpinning its success have been far from being implemented in many U.S. electricity markets. These have included vibrant retail competition, reduced concentration of generation ownership, and state and federal policies that encourage the development of adequate generation and transmission resources.

The leading "top-down" alternative for scarcity pricing has involved administrative intervention during shortage conditions that are typically reflected by emergency states and depletion of operating reserves (which could occur during summer peak hours when electricity use is near or at its annual peak). Such an approach was adopted at Midwest ISO, which would rely solely upon an administrative demand function for reserves to calculate an adder to the market-clearing price for energy when operating reserves were being drawn down to meet real-time demand. The "top-down" approach was well-suited at that point to the Midwest ISO wholesale market, given that none of the Midwest ISO states had either retail choice or a retail market nearly as active as ERCOT's.

Sole reliance on the "top-down" administrative pricing approach to produce scarcity pricing was, however, ruled out in ERCOT on the ground that it stood in conflict with the reliability needs of ERCOT's isolated interconnection and its active retail market. It was argued that in the absence of a must-offer provision for contracted resources, market participants would have an incentive to physically withhold generation to trigger scarcity pricing. Unlike in the Eastern or Western Interconnections, physical withholding of generation in ERCOT could cause severe reliability problems that would force the grid operator to intervene administratively to keep the lights on, potentially undermining the market. While a must-offer obligation is

present in markets with administrative scarcity pricing, in ERCOT it did not exist due to the retail market's need to avoid a regulatory requirement for bilateral contracting with a "must-offer" provision.

The Gordian knot of genuine scarcity pricing was dealt with by the PUCT staff in the classical way: by making a "decisive cut." The new rule gave small suppliers a safe harbor: If an entity were to control less than 5 percent of the installed capacity in ERCOT, it would be deemed not to have system-wide market power, and therefore would need not worry about prosecution if it decided not to offer any of its capacity into the market. On the other hand, exceeding the threshold would not necessarily mean the entity had market power. It would mean that if the supplier appeared to be withholding production or exercising economic withholding, and prices were being affected, the first question investigators would ask would be whether the entity had market power.

This "small fish swim free" approach was the result of an empirical review of balancing energy data by PUCT staff, which suggested that if the two or three largest generation fleets in ERCOT behaved as price-takers and allowed others to offer as they wished, the market would produce high prices when genuine scarcity resulted, as seen in the other single-settlement, energy-only markets of Alberta, Australia, and New Zealand. Sole reliance on high offers from market participants to set scarcity rents in the Texas market was still being questioned, however, and the debate was by no means over. In a report concerning a reliability event on March 3, 2008, when ERCOT experienced a large sudden drop in wind power, the ERCOT independent market monitor concluded that

> a) relying upon the submission of high-priced offers is an unreliable means of producing scarcity prices during scarcity conditions; and b) the price formation process during shortage conditions can become distorted if it does not include mechanisms to efficiently price the value of sacrificing the reserves that are required to maintain minimum reliability requirements.[38]

The quick disclosure of individual offer curves was another "bottom-up" feature of market power mitigation, in that it was intended to level the playing field and clarify when prices were a product of systematic market power

abuse or genuine scarcity—an alternative to the heavy, unit-specific mitigation seen in other U.S. markets.[39] Quick disclosure of individual offer curves ran against the grain of prevailing academic thinking, however, which suggested that revealed information on individual offers tacitly facilitated collusion among suppliers and helped them sustain high prices in excess of competitive levels. The literature concluded that the ability of competitors to cut prices "secretly" so that they were not exposed to retaliatory actions by their rivals was an essential element of competition.[40] These theoretical concerns had not been shown to be problematic in the Australian market, according to Australian market monitors who conducted an internal review of historical price data.[41] After careful consideration and consultation with several market power experts in academia and a number of wholesale power marketers, the PUCT staff finally decided to delay disclosure of various classes of information long enough to minimize any potential consequences of premature disclosure.[42]

ERCOT stakeholders, other market monitors, and even academics continued to express skepticism about this combination of the "small fish swim free" approach to system-wide market power and quick disclosure of resource-specific offer curves for energy and ancillary services, though the first two years of operation showed this "bottom-up" approach to be workable. Further evaluation of the effectiveness of the PUCT's approach would be possible after a few more years of operating experience.

Voluntary Mitigation Plan. In the ERCOT energy-only market, a supplier too large for the small-supplier exemption might also obtain advance protection against prosecution for market power abuse. Presumably, the supplier would submit a voluntary mitigation plan ensuring transparency in the availability of resources (to avoid physical withholding) and make the supplier a "price-taker" during times of scarcity (to avoid economic withholding). This safe harbor, however, would be specific to the supplier's own circumstances and subject to approval by the PUCT. The new rule allowed generators to apply for a voluntary mitigation plan that, if followed, would constitute an absolute defense against a finding of market power abuse with respect to the behaviors addressed in the plan. A large supplier could forgo the voluntary mitigation plan altogether if it believed it had no need for it.[43]

Challenges in the Transition to a Sustainable Energy-Only Approach

When the Australian energy-only resource-adequacy mechanism was introduced, competitive retailers had a contracting requirement for the first three years as a means to ensure grid reliability in the real-time market.[44] The offer cap, set at $A5,000 initially, also provided a very strong incentive for retailers to bring sufficient resources to cover their real-time positions. As such, the Australian energy-only resource adequacy mechanism appears to have worked well from its inception.

ERCOT, on the other hand, faced some challenges in making the transition to a sustainable energy-only approach. First, the energy-only resource-adequacy mechanism was added to the existing physical bilateral zonal market design, which had relied on out-of-market actions that affected real-time pricing in ways not consistent with the energy-only resource adequacy approved by the PUCT. Such actions included deployment of replacement reserves when the operator anticipated the depletion of the energy-balancing stack and deployment of spinning reserves for energy, or deployment of reliability must-run (RMR) resources to relieve intrazonal congestion and violation of voltage constraints. In the context of the newly designed nodal market, the deployment of incremental resources through the reliability unit commitment (RUC) process could be viewed as an out-of-market action.

Second, as mentioned above, competitive retailers in the ERCOT market had never had a contracting requirement that would guarantee sufficient resources being brought to the real-time market. Retailers and other load-serving entities likely experienced a learning curve in finding the correct balance between minimizing their procurement costs and reducing the risk of exposure to the real-time market.

Third, the gradual transition of an offer cap of $1,000 per MWh to $3,000 per MWh—levels far lower than the initial offer cap in the Australian market—had not, in some circumstances, provided sufficient incentive in the real-time market for quick-start, gas-fired generation to be available to respond to reliability events on the grid, and for loads to contract sufficiently to cover their load requirements in real time.

Responding to these circumstances, ERCOT stakeholders and the ERCOT independent market monitor (IMM) expressed concerns in 2007 that some of the defensive actions taken by ERCOT in fulfilling its reliability

mission, such as using conservative load forecast and procuring more responsive reserve service, were interfering with market prices and, in particular, with scarcity pricing. ERCOT operations staff acknowledged this problem and worked with stakeholders to address it. The "excessive" out-of-market actions by the system operator were attributed in part to sellers' poor performance of delivering the energy they promised under adverse weather conditions and selling more available capacity than they could provide in the real-time market. ERCOT tried to compensate for the reduced performance by "jumping the gun" in deploying nonspinning and replacement reserves, which resulted in suppression of market prices.

The chosen solution was to increase procurement of responsive reserves and impose stricter compliance standards, such as more frequent and random testing, on their providers. This would give ERCOT more headroom from a reliability standpoint, allowing its operators to move higher in the real-time energy bid stack than they had in 2007. The market would have a chance to clear near the top of the offer stack to set scarcity rents before ERCOT needed to take out-of-market actions.

The increase in the procurement of responsive reserves began on January 1, 2008. The ERCOT IMM reported to the ERCOT Technical Advisory Committee in early February that the chosen solution appeared to have sharply reduced the out-of-market actions by ERCOT operations and allowed the real-time energy market to function better.[45] This increase in the levels of responsive reserves could be considered an implicit mandatory insurance requirement (which is not much different from a capacity payment embedded in a bilateral contracting requirement), as loads would bear the uplifted cost for the additional resources. A parallel effort by ERCOT to improve reliability by enlisting demand-side participation through an emergency interruptible load service (EILS) was also implemented, and, as of this writing, discussion of further improvements along these lines was continuing.[46]

These actions should be considered temporary fixes that could be implemented quickly, given the prevailing software limitations and limited price-responsive demand available in the real-time market. Eventually, in a real competitive market, load response should set scarcity prices, and the PUCT has taken steps to realize that concept in the ERCOT market within the next few years.[47]

Will all of this help Texas reach the promised land of competitive electricity deregulation? ERCOT may have tight reserve margins over the next few years, so the energy-only approach will be put through its paces right out of the gate.[48] As of this writing, more than two years had passed since the decision by the commission to allow scarcity pricing in ERCOT spot markets, and on March 3, 2008, only two days after the cap was raised to $2,250 per MWh, balancing energy prices reached that level for three consecutive fifteen-minute intervals. In addition, many price spikes took place within the ERCOT balancing energy market during March–June 2008, resulting in five retail electric providers (REPs) failing to meet their financial obligations. These prices, which were far higher than allowed in any other North American markets, prompted a large public backlash and at least one legislative hearing.[49] While the trends look promising, however, the ERCOT market has yet to provide the levels of reliability by itself (that is, without the intervention of the ERCOT operator) that are seen in the Australian market. There is no doubt that the energy-only approach to resource adequacy in Texas is much closer to the economic gold standard of a commodity market than the capacity mechanisms, such as FCM and RPM, that are currently being used in the Eastern Interconnection. This mechanism should be considered only provisionally successful, however, in comparison to its Australian counterpart.

Conclusion: Why an Energy-Only Approach in Texas?

The framework adopted in August 2006 by the PUCT for market power and resource adequacy is unique in the United States. It establishes an energy-only resource-adequacy mechanism in the ERCOT market that raises the offer cap above the $1,000 per MWh prevailing in other North American electricity markets.[50] The rule increases the role of market forces in determining wholesale electricity prices and enhances the information available to market participants by dramatically increasing market transparency through prompt information disclosure.

So why was Texas the first U.S. market to develop an energy-only resource-adequacy mechanism? The authors believe that a number of circumstances that contributed to its development were, with respect to the United States, unique to Texas. Because the PUCT was the regulator over the wholesale market, retail market, and transmission grid in ERCOT, commission staff had the freedom to be creative in developing a combination of best practices with innovative adaptations that the PUCT could implement by rule. In contrast, the New England ISO's Forward Capacity Market was a complicated solution based on a multiparty compromise among state commissions, the New England ISO, and market participants through an administrative law proceeding.

The commission staff took the approach of looking for best practices, even when they were tried in non-U.S. markets and informed by contacts with market monitors in the United States, Canada, Australia, and the Pacific Basin, and with the academic and consulting community addressing resource adequacy issues worldwide. The staff also needed to deepen its understanding of the nodal market design issues that were concurrently being debated and draw insights from them that could be applied to the energy-only resource-adequacy design.

In looking around the globe, there seems to be a strong correlation among energy-only wholesale market design, generation-friendly transmission and generation-interconnection policies, and successful retail competition. Transmission policy and generation-interconnection policy set in the early days of wholesale market deregulation in Texas in the mid-1990s allowed a flood of potential generation projects to enter the ERCOT market, with enough transmission always available to deliver the power from generators to loads without heavy congestion.[51] The successful retail market created greater pressure from retailers and industrial customers to avoid implementing a "top-down" capacity resource-adequacy mechanism in ERCOT than was present in other U.S. markets.

Texas has taken a different approach to address resource adequacy by allowing higher electricity prices to improve the possibility of adequate recovery of investment while imposing a much higher level of market transparency compared to all other markets in the United States. In addition, adequate flexibility allows refinements, if desired. A few more years of operation will provide the empirical basis for further evaluation of the effectiveness of Texas's approach to addressing the complex issue of resource adequacy.