HAS THE TEXAS ELECTRICITY MARKET ACHIEVED TOO MUCH DEMAND RESPONSE?

Jay Zarnikau
Frontier Associates LLC
4131 Spicewood Springs Rd., Suite O-3
Austin, Texas 78731
Ph: 512/372-8778 x-103
jayz@frontierassoc.com

SUMMARY

Efforts to encourage demand response in the Electric Reliability Council of Texas (ERCOT) market have introduced technical and economic problems, prompting concerns that ERCOT may have *too much* demand response, at least in some markets. This paper describes the design of ERCOT markets and some steps taken to facilitate demand response. Some of the economic and operational problems prompted by load response are described. Various solutions and proposals presently under examination to resolve these problems are critiqued. It is argued that flaws in the market design are responsible for these problems. Instead of trying to limit demand response, efforts should instead be focused on re-designing the market to facilitate even greater levels of demand response.

BACKGROUND

The response of energy consumers or loads to price signals is widely viewed as crucial to the efficient operation of power systems. As noted by FERC: "Demand response is essential in competitive markets, to assure the efficient interaction of supply and demand, as a check on supplier and locational market power, and as an opportunity for choice by wholesale and end-use customers."[1] The United States Congress recently affirmed the importance of expanding demand response opportunities as a matter of national policy.[2] The National Association of Regulatory Utility Commissioners has called for regulatory commissions to accommodate demand-side resources and "remove any unnecessary barriers to customer responses to such wholesale market price signals."[3] As electricity markets are redesigned to facilitate wholesale and/or retail competition, stakeholders and policymakers are challenged with ensuring that consumers are presented with accurate price signals and the appropriate incentives to react to those prices.

---

[1] Federal Energy Regulatory Commission. *Working Paper on Standardized Transmission Service and Wholesale Electric Market Design*, March 15, 2002.
[2] *Energy Policy Act of 2005*, Section 1252(f).
[3] NARUC, *Resolution Regarding Equal Consideration of Demand and Supply Responses in Electricity Markets*, July 2000.

Yet, responses by retail energy consumers to price changes and the introduction of demand-side resources (e.g., interruptible loads) into markets initially designed for supply-side resources (e.g., power plants) pose certain technical and economic challenges. Some of these problems have become illuminated in the ERCOT market. Two problems are highlighted in this paper:

- An (alleged) over-supply of interruptible loads providing responsive reserves.
- The operational difficulties faced by system operators when short-term load forecasts (and generation needs) change as a result of consumer response to price changes.

Although some market participants maintain that these problems are the result of *too much* demand response, these problems may actually be traced to market design flaws which can be easily corrected. ERCOT has far too little, rather than too much, demand response.

OVERVIEW OF THE RESTRUCTURED ERCOT MARKET

About 85% of the electricity needs in the largest electricity-consuming state in the U.S. are satisfied through the intra-state ERCOT market. This electricity market has undergone gradual restructuring over the past decade to introduce greater competition in the wholesale and retail segments of the industry and to relax regulatory oversight. Senate Bill 373, enacted in 1995, required the Public Utility Commission of Texas (PUCT) to establish rules to foster wholesale competition and create an Independent System Operator (ISO) to ensure non-discriminatory transmission access, an equitable interconnection process for new generation capacity, and customer protection. In the summer of 1997, ERCOT became the first ISO in the U.S. Further reforms occurred as a result of Senate Bill 7 (enacted in 1999), which allowed customers of the investor-owned utilities within ERCOT to choose among various retail electric providers for a retail supply of electricity beginning on January 1, 2002. SB 7 also prompted the establishment of formal markets for ancillary services and balancing energy.

Within Texas reside a large number of industrial facilities involved in chemical production, petrochemicals, air separation, pulp and paper manufacturing, and steel production which can withstand short interruptions in their electricity supply with modest economic loss. Traditionally, these facilities were served through interruptible tariffs, which provided an electrical supply to the facility at a lower level of reliability in return for a discounted price. Consequently, the state has a very large base of industrial facilities which can reduce or curtail electricity purchase in response to either an instruction from the ISO or in response to a price signal.

Prior to the full-scale restructuring of the ERCOT market in January 2002, ERCOT relied upon roughly 3,500 MW of interruptible load, group load curtailment programs, direct load control, and other load management programs to maintain reliability. As the ERCOT market underwent redesign in 1999 to 2001 to foster competition, the PUCT ordered ERCOT to "Develop new measures and refine existing measures to enable load resources a greater opportunity to participate in the ERCOT market." (PUCT, 2000).

The restructured Texas (ERCOT) market has been as successful as any restructured market in promoting demand response in a few limited areas. As discussed below, caps on the participation of interruptible loads providing responsive reserves have been reached. Many industrial energy consumers meet all or a part of their energy requirements with balancing energy (essentially, spot market power) and respond to prices that change every 15 minutes. Yet, ERCOT has the smallest overall "existing demand response resource contribution" of any market in the U.S., at about 3% of peak demand. (FERC, 2006, p. 87.)

USING INTERRUPTIBLE LOADS AS A RESPONSIVE RESERVE

Many loads served under interruptible tariffs prior to restructuring are now providing ancillary services to the market. The design of ERCOT's wholesale market permits Loads Acting as Resources or "LaaRs" to compete "head-to-head" against generation resources to provide ancillary services, such as responsive reserves (provided by interruptible loads with under-frequency relays and which also agree to manual deployment or curtailment within ten minutes of notice) and non-spinning reserves (which can be interrupted by the ISO with 30 minutes of notice).[4] LaaRs selected to provide this service through ERCOT's formal day-ahead market for ancillary services receive the market-clearing price. Alternatively, LaaRs may be self-arranged by a load-serving entity, in which case they would receive a negotiated price. The load levels of LaaRs providing responsive reserves are monitored by ERCOT every two or three seconds to verify their availability.

In light of the pre-restructuring levels of industrial load served under instantaneous interruptible tariffs and armed with under-frequency relays, it was recognized that as much as 2,000 MW of load might be interested and capable of providing responsive reserves. As ancillary services markets were designed, this raised two concerns among the ISO's system operators and ERCOT's Reliability and Operations Subcommittee:

- If too large a share of responsive reserve requirements were provided by LaaRs, then there might not be adequate generation resources providing responsive reserves. Generating units with governors are better able to stabilize frequency in response to small deviations in frequency than LaaRs with their *off-or-on*, *all-or-nothing* response.
- There may be a possibility of "over-shoot" situations (where too much interruptible load might trip-off at the same time and raise frequency to an unacceptably high level).

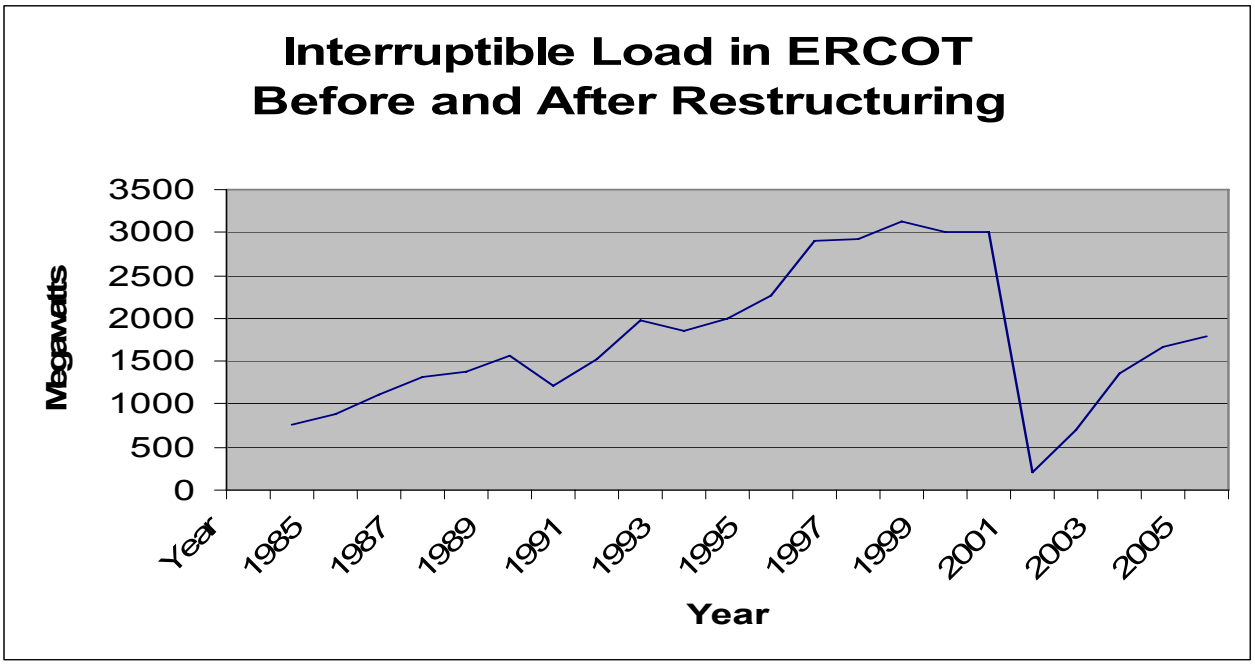
Consequently, limits were placed on the amounts of responsive reserves provided by LaaRs. Initially, this limit was 25% of ERCOT's requirements for this ancillary service (i.e., 575 MW each hour, given ERCOT's normal requirement of 2,300 MW). Later this was raised to 50% of ERCOT's need for responsive reserves (normally, 1,150 MW per

[4] LaaRs are permitted to provide Regulation on a pilot basis. In theory, LaaRs can also provide replacement capacity, although the systems necessary to permit interruptible loads to provide these services have not yet been fully implemented.

hour), as concerns surrounding over-shoot abated.[5] The presence of this cap frequently results in situations where a higher-priced generation resource is chosen to provide responsive reserves, despite the availability of a lower-priced demand-side resource. In addition, strict qualification criteria were introduced which precluded energy consumers whose load level could not be accurately predicted on a day-ahead basis from providing responsive reserves.

Within a couple years after LaaRs were permitted to participate in wholesale market, this cap was reached, as indicated in Figure 1. Currently 94 LaaRs (working with 10 scheduling entities) are qualified to provide ancillary services for a total capacity of 1,826 MW. Thus the amount of load qualified to provide responsive reserves is far in excess of the constraint on LaaR participation in this market (normally, 1,150 MW).

Figure 1



Sources: Data for 1985-1993 is from PUCT 1996 Statewide Electrical Energy Plan for Texas, June 1996, and represents interruptible loads plus a small contribution from various load cycling programs. Interruptible load data for 1994-1999 is from Project 22209 Annual Update of Generating Electric Utility Data, 2000. Data for 2000 and 2001 is from ERCOT Capacity, Demand and Reserve reports for those years. Data for 2002-2006 came from "Load Participation in ERCOT Ancillary Services Markets", April 18, 2006, AESP Brown Bag Seminar by Steve Krein, ERCOT staff.

LaaRs have been instrumental in preserving reliability, and typically one to three interruptions occur per year in response to under-frequency events. Competition posed by these LaaRs has served to reduce the market price of responsive reserves. While there

[5] ERCOT Staff for the Reliability and Operations Subcommittee, "Utilizing High-Set Load Shedding Schemes to Provide Response Reserve Services," November 2002.

has been a good mix of LaaRs by size, it is noteworthy that about one-half of the total quantity of LaaRs is provided by five very large industrial loads, as noted in Table 1.

Table 1: Categorization of LaaRs by Size

LaaR Capacity Range	Number of LaaRs	Total Capacity (MW)
1 to 10 MWs	66	283
11 to 50 MWs	20	388
51 to 100 MW	3	185
Greater than 100 MW	5	970

Source: ERCOT staff.

The excess supply of LaaRs relative to the cap has led to problems. As competition among LaaRs intensified for their limited share of the market for responsive reserves, many LaaRs began offering their interruption capability at increasingly-negative prices in hopes of securing a place among selected resources within the bid stack, and in anticipation that a higher-price generation resource would set the market-clearing price which all selected resources receive. (See Figure 2.) However, concerns emerged over the consequences of a market price as low as -$19,155 per MW! In such a case, all of the selected resources would then have to *pay* the market and the total costs could be substantial. This credit risk led to the imposition of a temporary floor price that prohibited negative bids. The current prevailing proposal is to continue to prohibit negative bids by LaaRs until ERCOT adopts a nodal structure in 2009. At that time, separate markets may be established for LaaRs and generators providing responsive reserves.

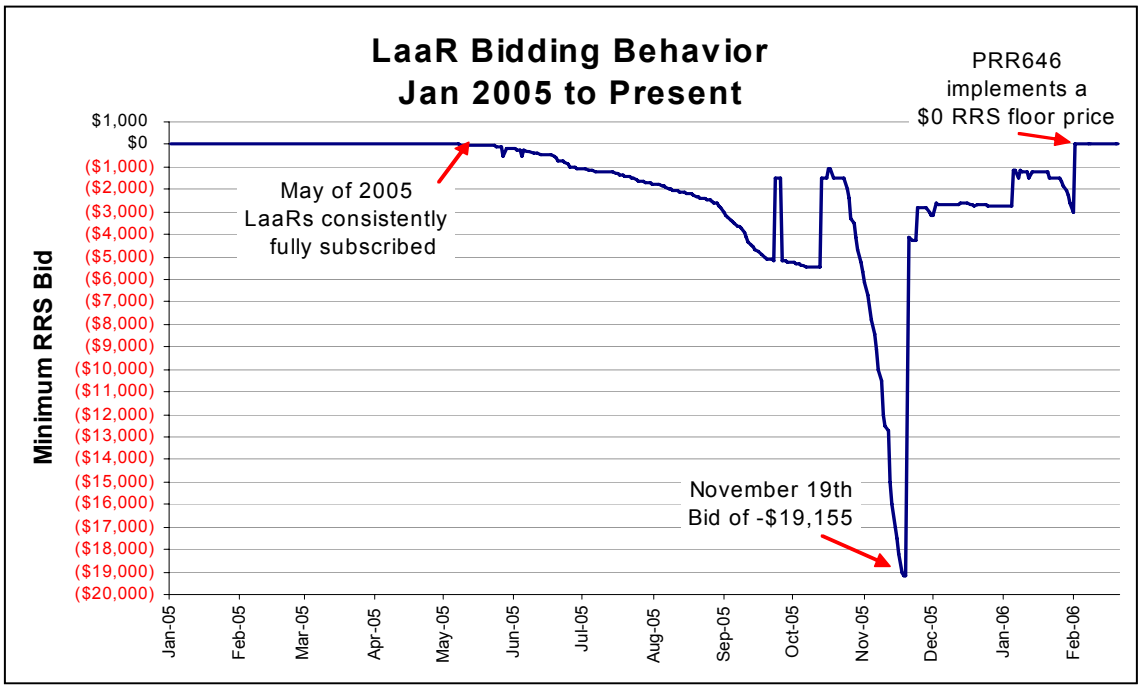
The cap and restrictive qualification requirements have also led to situations where interruptible loads have been ready, willing, and able to interrupt in order to balance supply and demand during reliability problems (e.g., the rolling blackouts which occurred on April 17, 2006), but could not be deployed since they were not selected by ERCOT to provide an ancillary service at that time.

Clearly, ERCOT has not yet adopted a policy which can fully utilize its base of load which can be instantaneously interrupted. Current debates center over the following proposals:

Increase the overall amounts of responsive reserves procured by ERCOT. As a result of the April blackouts, ERCOT started to bias its near-term load forecasts upward by one standard deviation of the historical error and procure expensive replacement capacity to ensure a greater cushion of operating reserves. This practice has led to excessive costs and some perverse market incentives. It is likely that a larger operating cushion could

more efficiently and equitably be achieved through the procurement of greater amounts of responsive reserves.

Figure 2



Source: ERCOT staff.

Establish a new ancillary service that is better tailored to take advantage of the characteristics of interruptible load. A proposal for a Tiered Frequency Response program has been offered by the Steel Mill Coalition. Under this proposal, interruptible loads that were not already providing an ancillary service would agree to interrupt during system emergencies and whenever system frequency dipped to 59.8 Hz or 59.75 Hz. Thus, these loads would be interrupted before LaaRs with under-frequency relays set at 59.7 Hz would be deployed. This service would be procured by the ISO through year-long contracts, rather than through day-ahead markets.

Require that LaaRs interrupt during an emergency, even if they are not providing an ancillary service at that time. This would increase the amount of demand reduction that could be achieved during an emergency, but would do little to tap the value of the interruptibility of loads that cannot meet the qualification standards to provide responsive reserves. It would also do nothing to address some of ERCOT's present difficulties in maintaining frequency within a tolerable range. Nonetheless, this proposal would provide some value at little, if no, cost to the market.

LOAD RESPONSE TO PRICE SIGNALS OR REP PROGRAMS

ERCOT's market structure provides some incentives for large energy consumers to reduce power purchases during peak or high-price periods, including:

- The design of transmission charges, which are based upon the consumer's contribution to demand during four summer month peaks.
- The ability to purchase balancing energy (through a retail electric provider or REP) and the design of the wholesale market settlements system rewards qualified scheduling entity (QSE) who can reduce generation needs during high price periods.
- Participation in demand response programs sponsored by REPs.

A large industrial energy consumer's transmission charge is based upon the consumer's contribution to ERCOT's coincident peak demand in four summer months. Often, transmission charges are treated as "pass-through" costs in the contracts offered by REPs. Consequently, larger energy consumers may see direct benefits by reducing their consumption during the four summer peaks, which are used to allocate transmission costs to consumers and REPs.

During the initial years of the restructured market (e.g., 2002 through 2005) consumers were free to deviate from scheduled load levels (in response to price changes, for example) with minimal penalties. "Passive load response" refers to a customer's deviation from its scheduled or anticipated load level in response to price signals (e.g., balancing energy prices or peak demand periods used to assign transmission costs) in situations where the customer has not formally offered this response to the market as a "resource." If the actual load level of a QSE turns out to be lower than its scheduled load level during a given 15-minute interval while its actual generation is equal to its scheduled generation, then the QSE is entitled to a payment or credit based on the energy imbalance multiplied by the balancing energy market price. This may provide energy consumers with an incentive to respond to wholesale market prices, provided their REP agrees to settle the consumer separately from other loads served through the REP. This separate settlement normally requires the metering of a load's consumption at 15 minute intervals. At market-open in 2002, interval data recorders (IDRs) were required on energy consumers with a billing demand over 1 MW. The IDR threshold was later reduced to 700 kW. How and whether an industrial energy consumer is compensated or credited for its response to wholesale prices and periods of peak demand is a contractual matter between the customer and its REP.

Some industrial energy consumers rely on balancing energy (essentially, spot market power) to meet some or all of their electricity needs, actively monitor 15-minute balancing energy prices, and reduce electricity purchases when prices exceed threshold levels. Whether QSE's are required to schedule the full quantity of their anticipated generation requirements with bilateral contracts with generation companies has been subject to changing policies. During the first years of the restructured market, there was a "balanced schedule requirement" (although some load serving entities ignored it). Later, a "relaxed balanced schedule requirement" was introduced, in part to encourage

REPs and large loads to rely in part on balancing energy to provide near-real-time price signals and foster demand response. Under the relaxed balanced schedule policy, a load serving entity could elect to purchase a share of its generation requirements from the balancing energy market. While this leaves the REP un-hedged and exposed to price fluctuations in the balancing energy market, many REPs found this strategy advantageous, particularly if they served loads with some capability to reduce energy usage in the face of high market prices. While volatile, balancing energy prices tend to be lower than the average cost of firm generation obtained from bilateral contracts. However, following the April 2006 blackouts, ERCOT's increased reliance upon replacement capacity and the practice of assigning the cost of procuring replacement capacity to QSEs who were "short" of scheduled capacity at the time the replacement capacity is needed greatly increased the cost of relying upon balancing energy to meet generation needs.

Finally, some industrial consumers of energy participate in curtailment programs that are established by their retail electric provider (REP). These are conducted outside of the formal ERCOT market and are used by REP's to shape their generation needs and reduce their costs.

If an energy consumer opts to offer its interruption capability into an ancillary services market, then its ability to react to wholesale balancing energy prices, avoid the four summer peaks, and participate in any REP-sponsored demand response programs will be constrained. If the load is providing responsive reserves, then ERCOT monitors the load's level every three seconds to ensure that the load is available for interruption should the system need to rely upon the interruption to maintain frequency. A QSE could incur a penalty (a scheduling control error) if it is not providing its committed level of operating reserves. Thus many of ERCOT's most flexible, interruptible, or potentially price elastic electric loads will not react to prices.[6]

The amount of load that is actively responding to price signals is difficult to quantify. Among the twenty largest industrial energy consumers in Houston, one or two are clearly responding to wholesale prices.[7] Total demand in ERCOT responds very little to wholesale price changes.[8]

The deviations of large loads from their forecasted or scheduled levels in response to wholesale balancing energy prices, avoid the four summer peaks, and participate in any REP-sponsored demand response programs has caused some scheduling and operating problems. The ERCOT staff has been unable to factor these responses into its short-term load forecasts[9] (although it is likely that bad weather forecasts have a much greater contribution to ERCOT's forecasting error than consumer response to price signals).

---

[6] This appears to be most true for LaaRs which have a bilateral contract to provide responsive reserves to a REP/QSE and are self-scheduled.

[7] Jay Zarnikau, Greg Landreth, Ian Hallett, and Subal Kumbhakar, "Industrial Energy Consumer Response to Wholesale Prices in the Restructured Texas Electricity Market," Draft, February 2005. Available at: http://www.frontierassoc.com/links.shtml.

[8] Analysis by Jay Zarnikau and Ian Hallett which will be presented in a forthcoming paper.

[9] ERCOT Staff report to the PUCT, Open Meeting of August 23, 2006.

These unexpected changes in demand complicate the ISO's challenge of matching supply and demand in real-time.

The solution approved by the PUCT involves completely removing any advanced notice of real-time prices when a nodal market structure is introduced in 2009. Also, a penalty (the Reliability Unit Commitment Capacity Short Charge) will be used to discourage REPs (and their customers) from relying upon the market (as opposed to bilateral contracts and the forthcoming day-ahead market) to secure generation. While the introduction of a day-ahead energy market may open up new demand response opportunities for loads that can accurately predict their energy requirements for each 5-minute or 15-minute period of the following day, loads without such foresight may be penalized. These changes will discourage loads from responding to price signals in real time unless price-responsive loads are shielded from such penalties. Further complications arise from the use of zonal prices to settle energy purchases, while zonal prices may be used to establish the value of a resource in the market. While taking these steps to discourage price-chasing may provide system operators with better demand forecasts, any demand response in real-time will be sacrificed.

As the PUCT approved "nodal protocols" which included these features to discourage demand response, it nonetheless expressed interest in exploring new avenues for demand response. Through a new project, the PUCT may soon explore mechanisms to:

- Insulate loads which are deemed to be price-sensitive from the Reliability Unit Commitment Short Charge.
- Provide some advance notice of real-time prices to loads (outside of the day-ahead market).
- Consider a proposal consistent with priority pricing, whereby loads would provide the ISO with a commitment to curtail at certain price points, in return for protection from various penalties and other incentives.

IN CONCLUSION

Other electricity markets would probably be envious of the size of the demand response *potential* available in the ERCOT market. Yet, ERCOT has not yet found a way to fully utilize this potential.

Attempts to accommodate a large base of instantaneously interruptible loads into a limited market for responsive reserves have created some problems. Passive load response and the presence of programs outside of the ISO's view have impaired the ability of system operators to match supply and demand in real-time.

Yet, each of these problems has simple solutions. New programs could be created to better exploit the potential value that can be provided by additional demand side resources. Over-shoot concerns can be addressed by establishing schemes to set under-frequency relays at different trip-off points. If system operators were provided with better information about likely responses by energy consumers to high balancing energy

prices or likely summer peaks, the ISO could factor that information into its near-term demand forecasts and reduce some of its present forecasting error. This demand elasticity information could be developed through better demand modeling, information from QSE's representing price-responsive loads, or through formal programs (e.g., a priority pricing program).

The need for vibrant demand response in ERCOT is only increasing. The PUCT is pursuing an "energy-only" resource adequacy mechanism, which places considerable emphasis on demand response to balance supply and demand over the long-run. ERCOT is probably the nation's most concentrated market, and present problems with supplier market power are likely to become exasperated under a nodal market structure.

ERCOT does not suffer from too much demand response. It instead suffers from a market design that cannot accommodate the full demand response potential.

REFERENCES

ERCOT Staff for the Reliability and Operations Subcommittee, "Utilizing High-Set Load Shedding Schemes to Provide Response Reserve Services," November 2002.

Energy Policy Act of 2005, Section 1252(f).

Federal Energy Regulatory Commission, Working Paper on Standardized Transmission Service and Wholesale Electric Market Design, March 15, 2002.

Federal Energy Regulatory Commission, *Assessment of Demand Response and Advanced Metering*, August 2006.

NARUC, *Resolution Regarding Equal Consideration of Demand and Supply Responses in Electricity Markets*, July 2000.

Public Utility Commission of Texas, *Final Order in Docket No. 23220: Petition of the Electric Reliability Council of Texas for Approval of the ERCOT Protocols*, 2000.

Zarnikau, Jay, Testimony in Docket No. 31540: Proceeding to Consider Protocols to Implement a Nodal Market in ERCOT Pursuant to PUC Subst. R. 25.501. November 2005.

Zarnikau, Jay, Greg Landreth, Ian Hallett, and Subal Kumbhakar, "Industrial Energy Consumer Response to Wholesale Prices in the Restructured Texas Electricity Market," Draft, February 2005. Available at: http://www.frontierassoc.com/links.shtml

Zarnikau, Jay, "Using Interruptible Load as an Ancillary Service in the Restructured ERCOT Market," *US Energy Association Dialogue*, July, 2006. Available at: http://www.usaee.org/pdf/Aug06.pdf#13d