POTOMAC ECONOMICS

2003 STATE OF THE MARKET REPORT
FOR THE
ERCOT WHOLESALE ELECTRICITY MARKETS

POTOMAC ECONOMICS, LTD.

Advisor to the Market Oversight Division
Public Utility Commission of Texas

August 2004

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (CONT.)

LIST OF TABLES

ACKNOWLEDGMENTS

We wish to acknowledge the helpful input and numerous comments provided by the staff of the Market Oversight Division of the Public Utility Commission of Texas, including Parviz Adib, Richard Greffe, Danielle Jaussaud, Julie Gauldin, and David Hurlbut. We are also grateful for the assistance of ERCOT in providing the data used in this report and in responding to our inquiries regarding the operation of the market.

EXECUTIVE SUMMARY

In this report we review a broad range of issues associated with the ERCOT wholesale electricity market in 2003. This includes summarizing the market outcomes, assessing the efficiency and incentives provided by the current market rules and procedures, and evaluating the conduct of market participants. Based on the results of these analyses, the report finds that the performance of the markets has been improving due to changes in the market rules and software, and experience gained by the market participants. However, the ERCOT market suffers from a number of operational issues and inefficiencies that can be attributed to the portfolio scheduling and bidding rules and procedures, and the zonal congestion management framework.

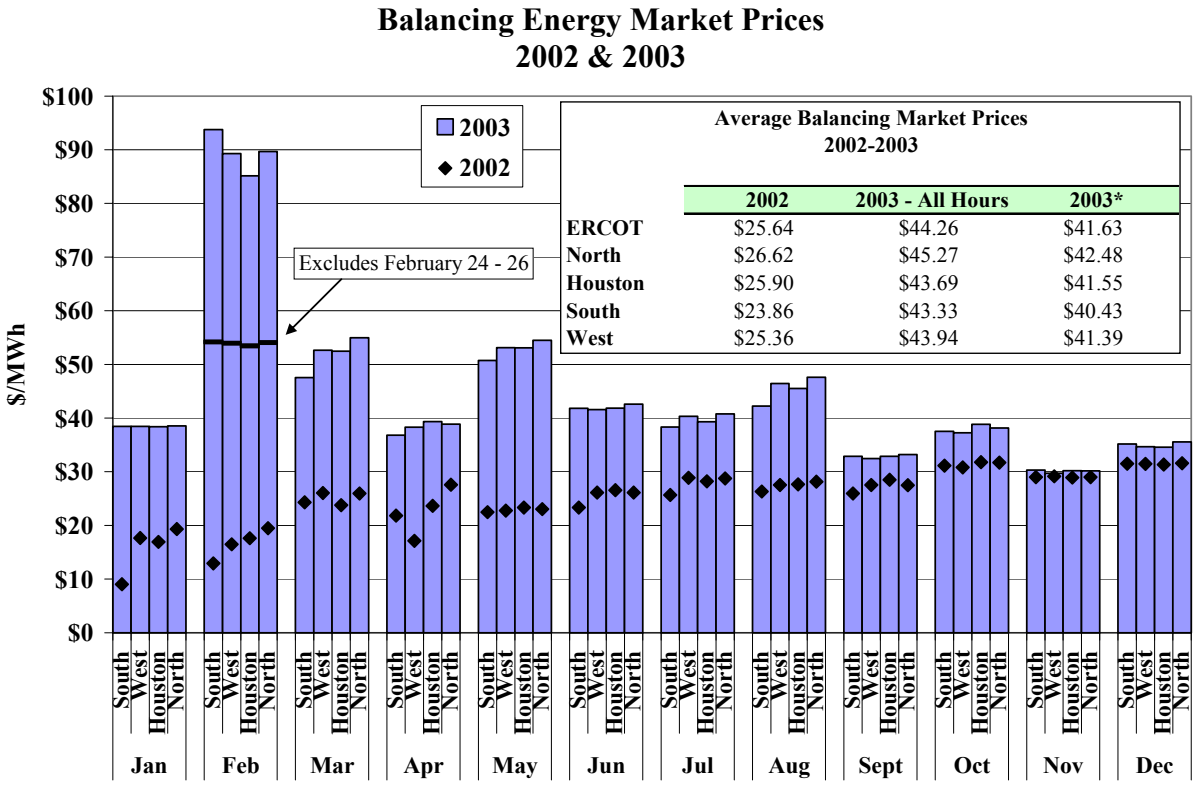
To address these issues, we make a number of recommendations designed to improve the performance of the current ERCOT markets. However, a number of the market issues identified in this report would be most effectively addressed by the introduction of the Texas Nodal markets that are currently being considered for implementation in 2006. In the sections below, we review the results and recommendations of the report in various areas, beginning with a review of the results of the balancing energy market and ancillary services markets in 2003.

A. Review of Market Outcomes

1. Balancing Energy Prices

The balancing energy market allows participants to make real-time purchases and sales of energy in addition to their forward schedules. While only a small portion of total electricity produced in ERCOT is cleared through the balancing energy market, its role is critical in the overall wholesale market.

The balancing energy market governs real-time dispatch of generation by altering where energy is produced in order to: a) manage interzonal congestion, and b) displace higher-cost energy with lower-cost energy given the energy offers of the Qualify Scheduling Entities ("QSEs"). In addition, the balancing energy prices also provide a vital signal of the value of power for market participants entering into forward contracts. Although most power is purchased through forward contracts of varying duration, the spot prices emerging from the balancing energy market should directly affect forward contract prices. The following figure shows the monthly average balancing energy prices in 2002 and 2003.



*Excludes February 24-26.

This figure shows that balancing energy prices in 2003 were substantially higher than in 2002. Load-weighted average prices throughout ERCOT were more than 70 percent higher in 2003 than in 2002. This was primarily due to considerable increases in natural gas prices. Average natural gas prices increased in 2003 by more than 65 percent from 2002 levels. Prices were substantially higher in most hours in 2003 relative to 2002, as higher fuel prices caused increased electricity prices over nearly the entire range of hours for the year.

The figure also shows that the highest prices during 2003 occurred in the spring. Only some of these high prices can be explained by increases in natural gas prices during this period. The high prices shown in February reflect price spikes that occurred in the balancing energy and ancillary service markets during the period of February 24 to 26 in response to unusually high loads associated with a period of extremely cold weather and a spike in natural gas prices. The PUCT Market Oversight Division issued a report on this event and concluded that the primary factors contributing to the price spikes on February 24 to 26: were (a) the inability of generators to procure sufficient natural gas on February 24; (b) the failure of many QSEs to provide ERCOT

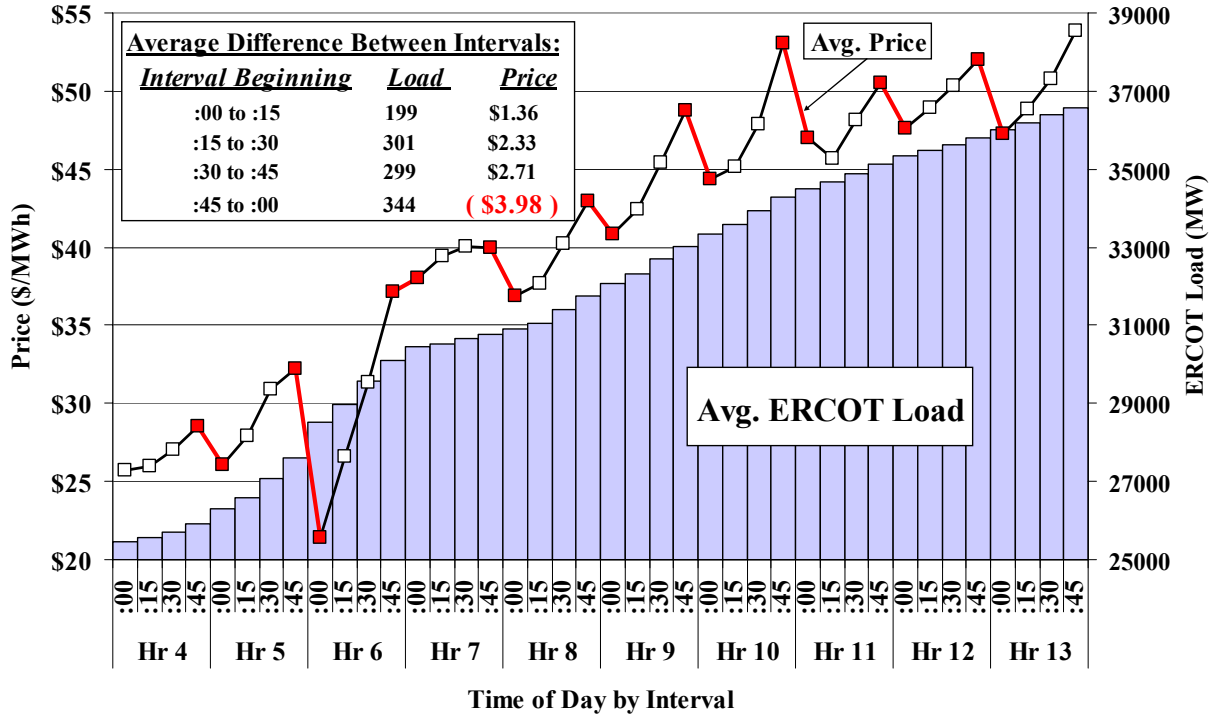
with timely updates on the status of specific generating units; and (c) "hockey stick" bidding by one QSE. We agree with these conclusions. Another important factor that influences the balancing energy market results, particularly under tight conditions like those in February 2003, is that less than half of the available energy is offered in the ERCOT balancing energy market on average. Based on the detailed assessment described later in this report, we conclude that the primary reason why such a considerable amount of available energy is not offered is attributable to the current market design. This raises concerns because it can cause very high-priced offers to be deployed and set high prices when no shortage exists.

Lastly, the figure above shows that balancing energy price differences between the zones within ERCOT were relatively small, reflecting moderate amounts of interzonal congestion. In both years, the North Zone exhibited the highest average prices and the South Zone exhibited the lowest. The average difference in prices between these two zones was approximately 7.0 percent in 2002 and 4.5 percent in 2003.

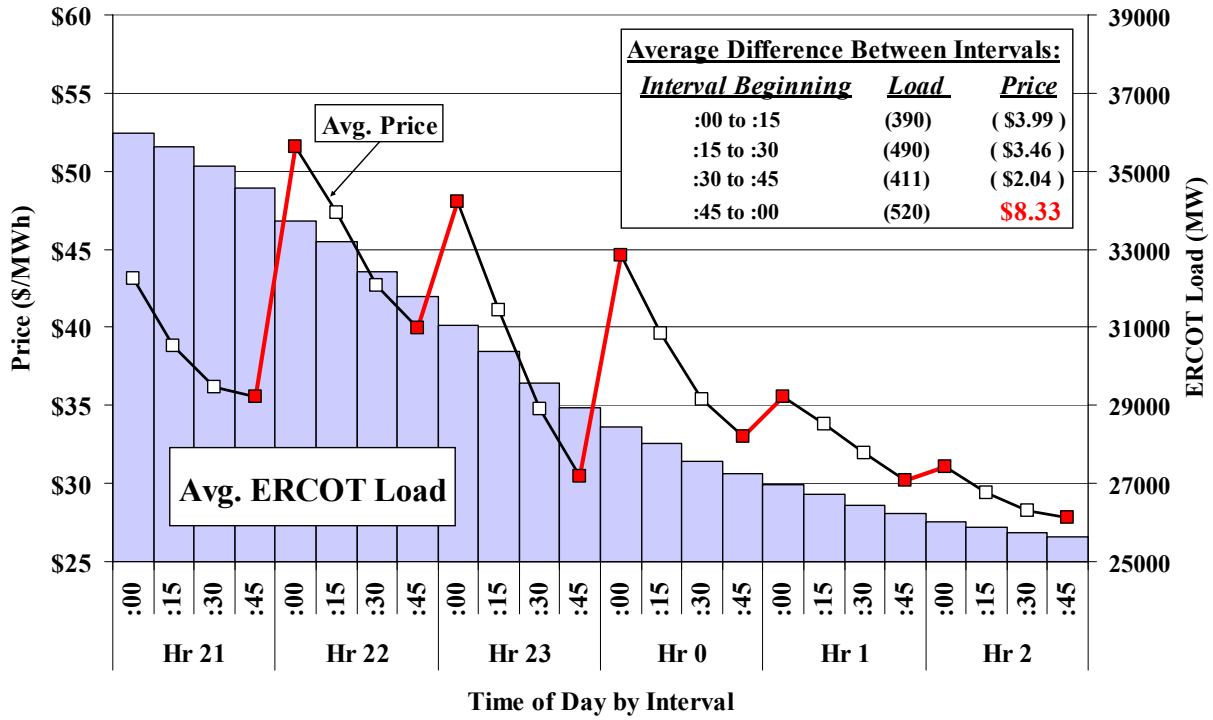
The report evaluates two other aspects of the balancing energy prices: 1) the primary determinants of the prices, and 2) the correlation of the prices with forward electricity prices in Texas. With regard to the determinants of balancing energy prices, one should expect that prices would be primarily determined by load levels and fuel prices in a well-functioning spot market. Although there is a strong relationship between fuel prices and balancing energy prices, we do not observe a strong relationship between prices and actual load levels in ERCOT. Instead, we observe a clear relationship between the net balancing energy deployments and the balancing energy prices, which is unexpected in a well-functioning market.

The report concludes that the observed relationship is primarily due to the hourly scheduling patterns of most of the market participants. We observe that the energy schedules change by large amounts at the top of each hour while load increases and decreases smoothly over time. This creates extraordinary demands on the balancing energy market and erratic balancing energy prices, particularly in the morning when loads are increasing rapidly and in the evening when loads are decreasing rapidly. The following figures summarize these erratic price patterns by showing the balancing energy prices and actual load in each 15-minute interval during the morning "ramping-up" hours and evening "ramping-down" hours.

Average Balancing Energy Prices and Load by Time of Day
Ramping-Up Hours -- 2003



Ramping-Down Hours – 2003



These pricing patterns and the fact that balancing energy prices are not as strongly correlated with actual load as expected raises significant efficiency concerns regarding the operation of the balancing energy market. These concerns and our recommendations to address these concerns are discussed below.

Finally, with regard to the correlation of the balancing energy prices and forward contract prices in Texas, the report finds improved correlation in 2003. A general principle that applies in the markets for most commodities is that spot markets and forward markets for the commodity will be closely correlated when barriers to arbitraging the two markets are low (firms can freely buy and sell in either market). Hence, sharp changes in the spot prices will be reflected in forward prices as well. In ERCOT, the spot market for electricity is the balancing energy market. The report shows that the correlation of the balancing energy prices with bilateral forward prices improved in 2003 from 2002. This improvement is likely due, in part, to the introduction of "relaxed balanced schedules" in November 2002, which increased participants' flexibility to arbitrage prices by allowing them to schedule more or less energy than their expected load. This reduced the barriers to participants buying or selling more actively in the balancing energy market.

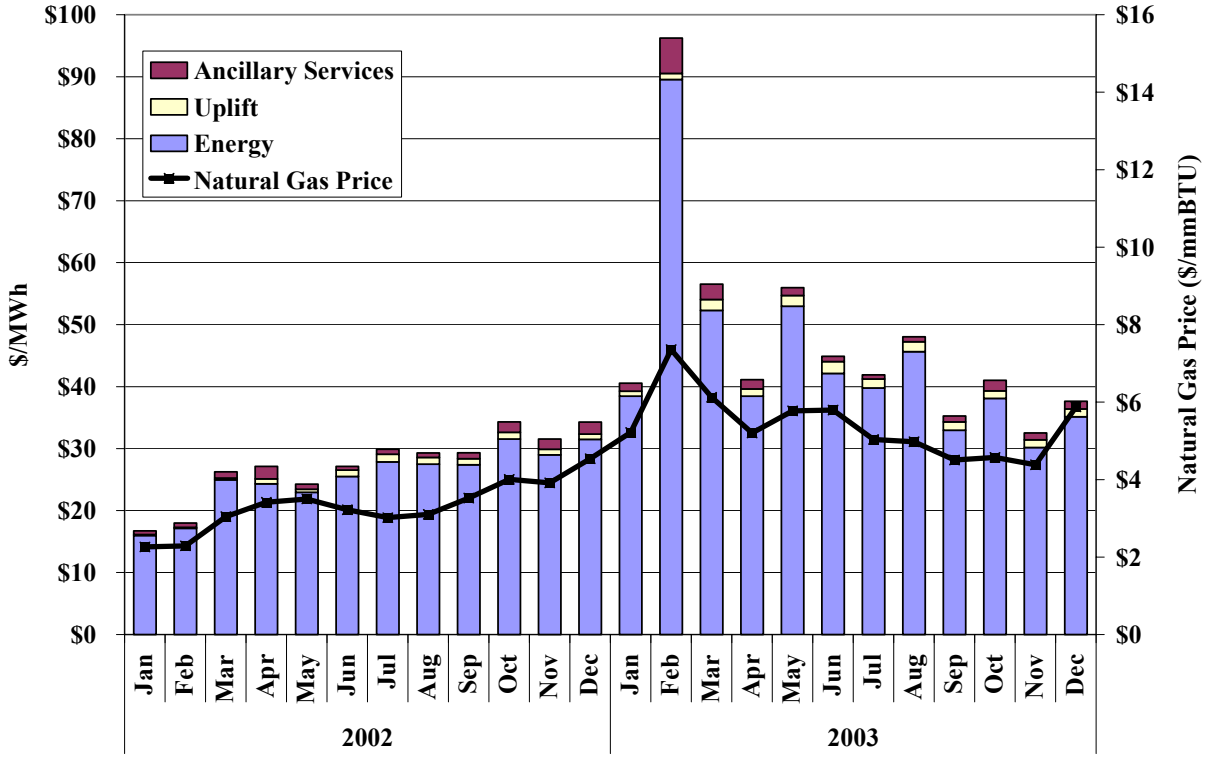
The report also shows that the volatility of the balancing energy prices increased in 2003 and remains substantially higher than the volatility of day-ahead bilateral prices. Assuming market participants are willing to pay a premium for protection from real-time price volatility, we would expect bilateral prices to be slightly higher than the balancing energy prices on average. This was not the case in 2003, although the average day-ahead prices were slightly higher than the average balancing energy prices when one excludes the extreme balancing energy prices during February 24-26.

2. All-In Electricity Prices

In addition to the costs of energy, loads incur costs associated with operating reserves, regulation, and uplift. The uplift costs include payments for out-of-merit capacity ("OOMC"), out-of-merit energy ("OOME"), and reliability must run agreements ("RMR"). We calculated an average all-in price of electricity that includes balancing energy costs, ancillary services costs, and uplift costs.

The monthly average all-in energy prices for the past two years are shown in the figure below along with a natural gas price trend. The all-in prices for electricity increased by 72 percent in 2003 from the 2002 levels. A substantial portion of this price increase can be explained by the significant increase in natural gas prices that occurred in 2003. In addition, the uplift associated with local congestion also increased considerably in 2003. These costs, regardless of the location of the congestion, are borne equally by all loads within ERCOT.

Average All-in Price for Electricity in ERCOT
2002 to 2003



The report compares the ERCOT all-in electricity prices to those of the four other major centralized wholesale markets and finds:

- In 2002, ERCOT exhibited the lowest all-in price, 25 percent lower than the next lowest-priced market.
- ERCOT also experienced the largest increase in prices between 2002 and 2003. This is not surprising given that ERCOT relies more heavily on natural gas-fired generators than any other region.
- Even with the increase in 2003, the all-in prices in the ERCOT region remain relatively low due in part to its substantial resource margin.

3. Ancillary Services Markets

ERCOT incurred higher costs for reserves and regulation than other markets in 2003. This is due in part to the higher quantities of regulation and responsive reserves that are required in ERCOT due to its limited interconnections with adjacent areas. This report concludes that the ancillary services prices in ERCOT are generally higher than expected. For example, responsive reserves prices averaged more than $11 per MWh, which is substantially higher than similar markets in other regions. We identify two explanations for this:

- A considerable portion of the available capability in ERCOT is not scheduled or offered in the ancillary services markets. Less than one-third of the regulation capability was scheduled or offered in the regulation market in 2003, while approximately 50 percent of the available responsive reserves capability and 25 percent of the non-spinning reserves capability were scheduled or offered.
- The sequential design of the ERCOT ancillary services and energy markets (ancillary services are procured in advance of the energy market rather than being jointly-optimized with the dispatch of energy) leads to higher costs because the allocation of resources to provide ancillary services will be suboptimal. The only market with higher responsive reserves prices is PJM, which also does not jointly-optimize the procurement of reserves and energy.

We understand that co-optimization is being contemplated in the design of the Texas Nodal markets that are currently under consideration. If the Texas Nodal markets are adopted, we would encourage implementation of ancillary services markets that are co-optimized with the nodal energy markets.

4. Net Revenue Results

A final analysis of the outcomes in the ERCOT markets in 2003 is the analysis of "net revenue". Net revenue is defined as the total revenue that can be earned by a new generating unit less its variable production costs. It represents the revenue that is available to recover a unit's fixed and capital costs and reflects the economic signals provided by the market for investors to build new generation or for existing owners to retire generation. In long-run equilibrium, the markets should provide sufficient net revenue to allow an investor to break-even on an investment in a new generating unit.

In the short-run, if the net revenues produced by the market are not sufficient to justify entry, then one of three conditions likely exists:

(i) New capacity is not currently needed because there is sufficient generation already available;

(ii) Load levels, and thus energy prices, are temporarily below long-run expected levels due to mild weather or economic conditions; or

(iii) Market rules are causing revenues to be reduced inefficiently.

Likewise, the opposite would be true if prices provide excessive net revenue in the short-run. Excessive net revenue that persists for an extended period in the presence of a capacity surplus is an indication of competitive issues or market design flaws.

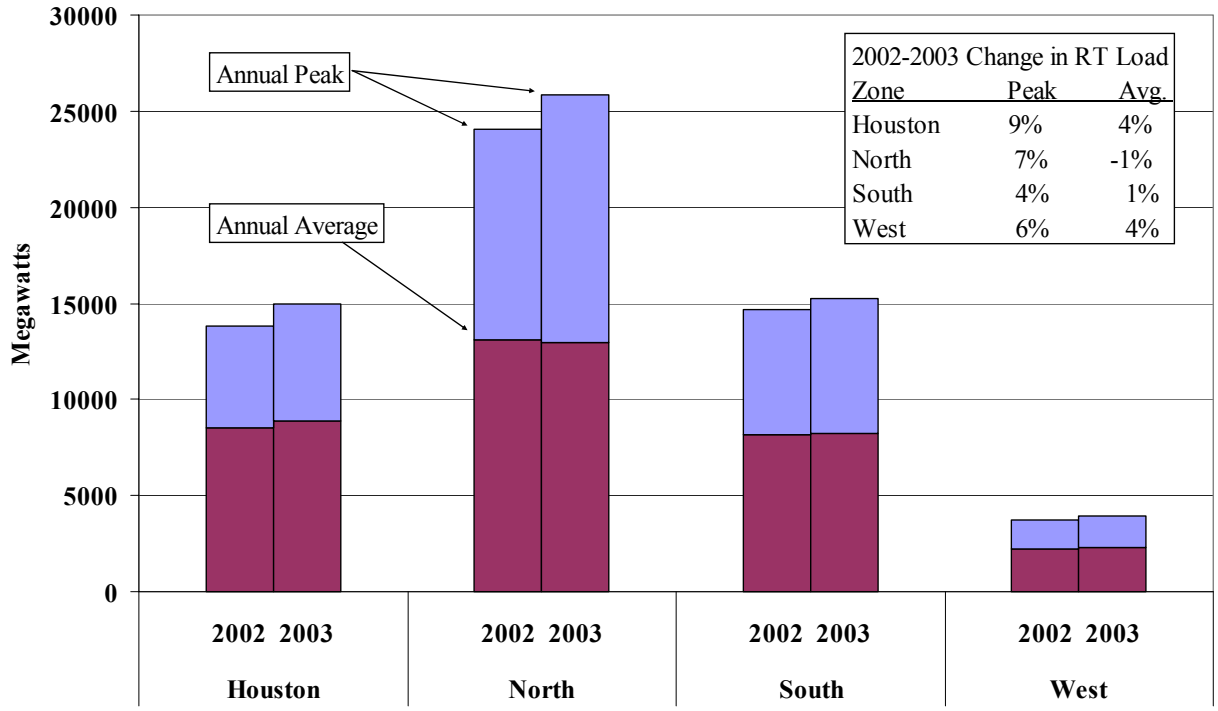
The report estimates the net revenue that would have been received in 2002 and 2003 for two types of units, a natural gas combined-cycle and a natural gas turbine. The net revenue increased significantly in 2003 from 2002, largely due to higher natural gas prices and increased balancing energy purchases in 2003, both of which led to higher balancing energy prices. Despite this rise in net revenue in 2003, neither type of new generating unit would have earned sufficient net revenue to make the investment profitable. This is not surprising given the surplus of capacity that currently exists in ERCOT.

B. Demand and Resource Adequacy

1. Electrical Loads in 2003

Load levels remain one of the fundamental factors that determine the conditions in any electricity market. Because electricity cannot be stored, the electricity market must ensure that generation matches load on a continuous basis. The figure below shows that load increased on average by only 1.4 percent from 2002 to 2003. However, the increase during peak demand conditions was almost 7 percent due to relatively hot weather during the summer of 2003.

Annual ERCOT Load Statistics by Zone



Significant changes in these peak demand levels are very important because they determine the probability and frequency of shortage conditions, although no shortages occurred under peak demand conditions in 2003 due to ERCOT's relatively high resource margins. More broadly, peak demand levels and capability of the transmission network are the primary factors that determine whether the existing generating resources are adequate to maintain reliability.

2. Generation Capacity In ERCOT

The report also provides an accounting of the current ERCOT generating capacity, which is dominated by natural gas-fired resources. These resources account for 73 percent of generation capacity in ERCOT as a whole, and 85 percent in the Houston Zone. This makes ERCOT particularly vulnerable to natural gas price spikes because the other resource types (coal and nuclear) are primarily base load units. There is approximately 20,000 MW of coal and nuclear generation in ERCOT. Because there were only 23 hours during 2003 when ERCOT load was less than 20,000 MW, it was necessary for natural gas resources to be running in most hours.

Our analysis also shows that ERCOT has substantial excess capacity. Resource margins (the percentage by which total capacity exceeds peak demand) for ERCOT as a whole have increased

by two to five percent in 2003, depending on types of capacity included in the calculation. When import capability, resources that can be switched to the SPP, and Loads acting as Resources are excluded from the calculation, the resource margin in 2003 was 24 percent. When these classes of capacity are included, the resource margin is 33 percent. This is particularly notable given that the 2003 peak demand level was approximately 2 GW higher than the planning forecast level. At the zonal level, resource margins are above 20 percent for all four zones and are as high as 37 percent in the South Zone and 45 percent in the West Zone.

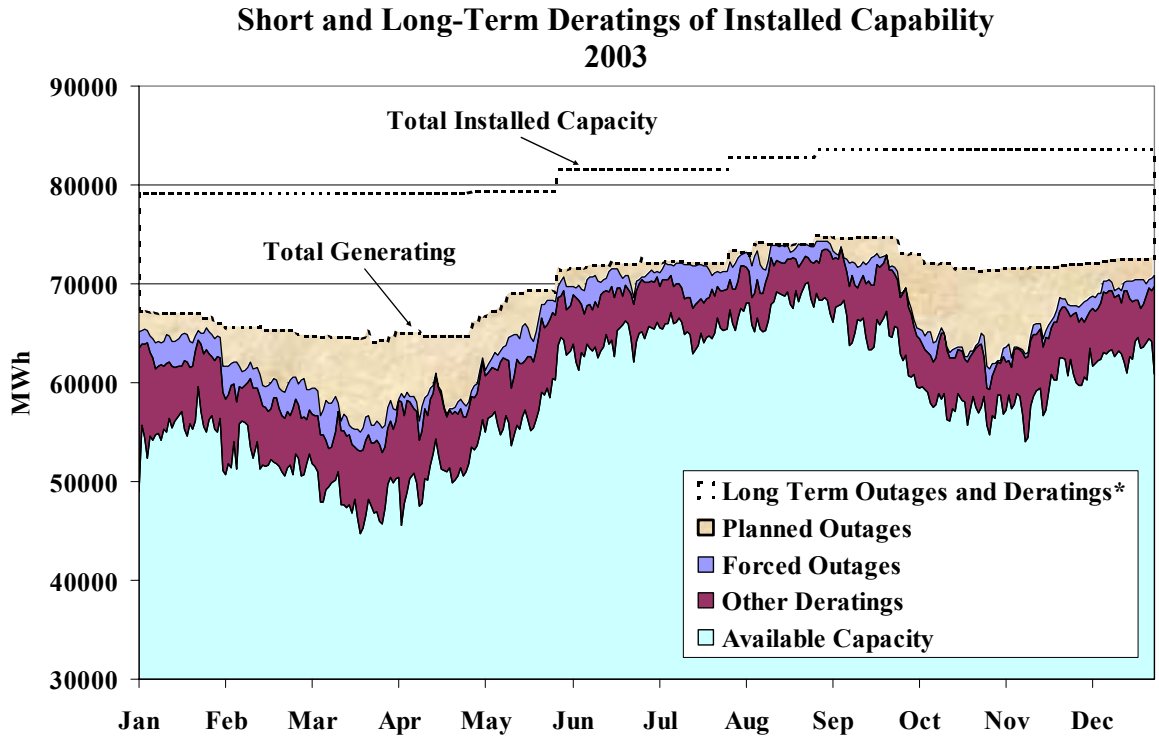
Although these resource margins are sizable, it is important to consider that electricity demand in Texas has been growing at a rapid pace. From 1994 to 2003 the coincident peak grew at an annual rate of 3.6 percent,[1] despite a significant decline in 2001 due to the recession. At this rate it will take little more than three years to reduce the ERCOT resource margin to 15 percent. It is also important to consider that a significant number of generating units in Texas are soon reaching or are already exceeding their expected lifetimes.

3. Generator Outages and Commitments

Despite the relatively high resource margins, resource adequacy must be evaluated in light of the resources that are actually available on a daily basis to satisfy the energy and operating reserve requirements in ERCOT. A substantial portion of the installed capability is frequently unavailable due to generator deratings.

A derating is the difference between a generating resource's installed capability and its maximum capability (or "rating") in a given hour. Generators can be fully derated (rating equals 0) due to a forced or planned outage. However, it is very common for a generator to be partially derated (e.g., by 5 to 10 percent) because the resource cannot achieve its installed capability level due to technical or environmental factors (e.g., ambient temperature conditions). The following figure shows the daily available and derated capability of generation in ERCOT.

[1] ERCOT Transmission Study, 2003, p. 56.



This figure shows that long-term outages and deratings typically exceed 10 GW and were as high as 14 GW in the spring and fall of 2003. These long-term deratings reduce the effective resource margins in ERCOT from the levels reported above. Most of these deratings reflect:

- Resources out-of-service for extended periods due to maintenance requirements;
- Resources out-of-service for economic reasons (e.g., mothballed units); or
- Output ranges on available generating resources that are not capable of producing up to the full installed capability level.

With regard to short-term deratings and outages, the patterns of planned outages and forced outages were consistent with expectations:

- Forced outages occurred randomly over the year and the forced outage rates were relatively low (although all forced outages may not be reported to ERCOT).
- Planned outages were relatively large in the spring and fall and extremely small during the summer, as expected.

The "other deratings" shown in the figure ranged from an average of 5 percent during the summer in 2003 to as high as 10 percent in other months. These deratings include outages not

reported or correctly logged by ERCOT and natural deratings due to ambient conditions and other factors. These outages and deratings do not raise any significant issues.

In addition to the generation outages and deratings, the report evaluates the results of the generator commitment process in ERCOT, which is decentralized and largely the responsibility of the QSEs. This evaluation includes analysis of the real-time excess capacity in ERCOT. We define excess capacity as the total online capacity plus quick-start units each day minus the daily peak demand for energy, operating reserves, and up regulation. Hence, it measures the total generation available for dispatch in excess of the electricity needs each day.

The report finds that excess capacity is significant in ERCOT, averaging almost 9,000 MW and rarely falling below 5,000 MW. These results show that the ERCOT system is generally over-committed, indicating significant inefficiencies in the outcomes of the current ERCOT markets. The tendency to over-commit capacity can be attributed in large part to the lack of a centralized day-ahead commitment process in ERCOT. Without a centralized commitment mechanism, each participant makes independent generator commitment decisions that, taken together, are not likely to be optimal. Hence, the introduction of day-ahead energy and operating reserves markets under the Texas Nodal market design currently being considered promises substantial efficiency improvements in the commitment of generating resources.

4. Load Participation in the ERCOT Markets

The ERCOT Protocols allow for loads to participate in the ERCOT-administered markets as either Load acting as Resources ("LaaRs") or Balancing Up Loads ("BULs"). LaaRs are loads that are qualified by ERCOT to offer responsive reserves, non-spinning reserves, or regulation into the day-ahead ancillary services markets and can also offer blocks of energy in the balancing energy market.

There are 35 participants qualified as LaaRs with a total capability of 1,200 MW. In 2003, LaaRs provided an average of 800 MW of responsive reserves to the market, divided equally between self-supplied reserves and offers cleared in the responsive reserves market. This represents a relatively large share of the total 2,300 MW requirement for responsive reserves. Under ERCOT rules, LaaRs were permitted to supply up to 35 percent (approximately 800 MW) of the responsive reserves requirement in 2003. ERCOT rules have recently been changed to

allow LaaRs to supply up to 1,150 MW of the requirement. Although these 35 participants are qualified to participate in non-spinning reserves and balancing up energy markets, the only market they have selected to participate in so far is the responsive reserves market. LaaRs have actually been called to reduce consumption through the automated activation under frequency relays twice in 2003.

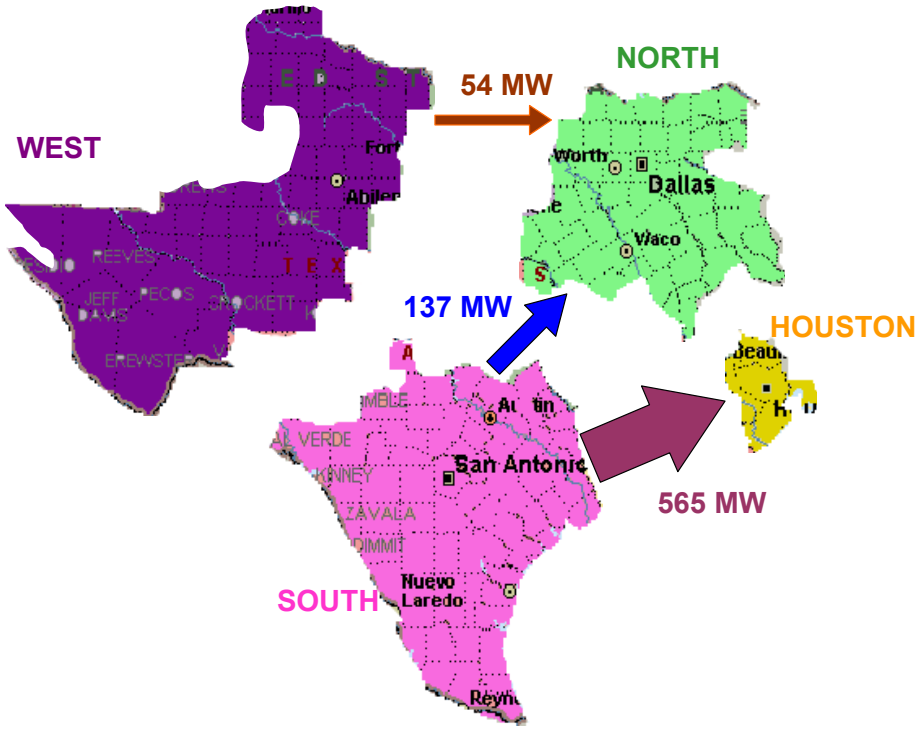
C. Transmission and Congestion

One of the most important functions of any electricity market is to manage the flows of power over the transmission network, limiting additional power flows over transmission facilities when they reach their operating limits. In ERCOT, constraints on the transmission network are managed in two ways. First, ERCOT is made up of zones with the constraints between the zones managed through the balancing energy market. The balancing energy market increases energy production in one zone and reduces it in another zone to manage the flows between the two zones when the interface constraint is binding (i.e., when there is interzonal congestion). Second, constraints within each zone (i.e., local congestion) are managed through the redispatch of individual generating resources. The report evaluates the ERCOT transmission system usage and analyzes the costs and frequency of transmission congestion.

1. Electricity Flows between Zones

The balancing energy market uses the Scheduling, Pricing, and Dispatch ("SPD") software that dispatches energy in each zone in order to serve load and manage congestion between zones. The SPD model embodies the market rules and requirements documented in the ERCOT protocols. To manage interzonal congestion, SPD uses a simplified network model with four zone-based locations and three transmission interfaces. The three transmission interfaces are referred to as Commercially Significant Constraints ("CSCs"). These are the West to North interface, the South to North interface, and the South to Houston interface. The following figure shows the average flows modeled in SPD during 2003 over each of these CSCs.

Average Modeled Flows on Commercially Significant Constraints
2003



The analysis of these CSC flows in this report indicates that:

- The average flow across the CSCs as modeled by SPD is relatively low, with the exception of the South to Houston flow that averages 565 MW. This shows the load in each zone is modeled as being primarily satisfied by resources within the zone rather than imports from other zones.
- The simplifying assumptions made in the SPD model can result in modeled flows that are considerably different from actual flows.
- A considerable quantity of flows between zones occurs over transmission facilities that are not defined as part of the three primary CSCs. When these flows cause congestion, it is beneficial to create a new CSC, such as the North to Houston CSC implemented by ERCOT in 2004, to better manage congestion over that path.

2. Interzonal Congestion Costs and Congestion Rights

When interzonal congestion arises, higher-cost energy must be produced within the constrained zone because lower-cost energy cannot be delivered over the constrained interfaces. When this occurs, participants must compete to use the available transfer capability between zones. In order to allocate this capability in the most efficient manner possible, ERCOT establishes a

clearing price for each zone and the price difference between zones is charged for any interzonal transactions. The levels of interzonal congestion were modest in 2003, totaling approximately $25 million.

Participants in Texas can hedge against congestion in the balancing energy market by acquiring Transmission Congestion Rights ("TCRs") between zones which entitle the holder to payments equal to the difference in zonal balancing energy prices. The report shows that the quantity of TCRs defined over a congested CSC frequently exceeds the total modeled flows over the CSC. When this occurs, the congestion revenue collected by ERCOT will be insufficient to satisfy the financial obligation to the holders of the TCRs and the revenue shortfall is collected from loads through uplift charges.

These issues will be studied in more detail in the market operations report to be issued this fall. However, improvements to the existing zonal markets in these areas are expected with the implementation of a Protocol revision that will reduce the percentage of Annual TCRs and increase the percentage of monthly TCRs sold to market participants. Ultimately, the introduction of the Texas Nodal markets currently under consideration would fully address these issues by eliminating the inconsistencies between the market software and the actual system.

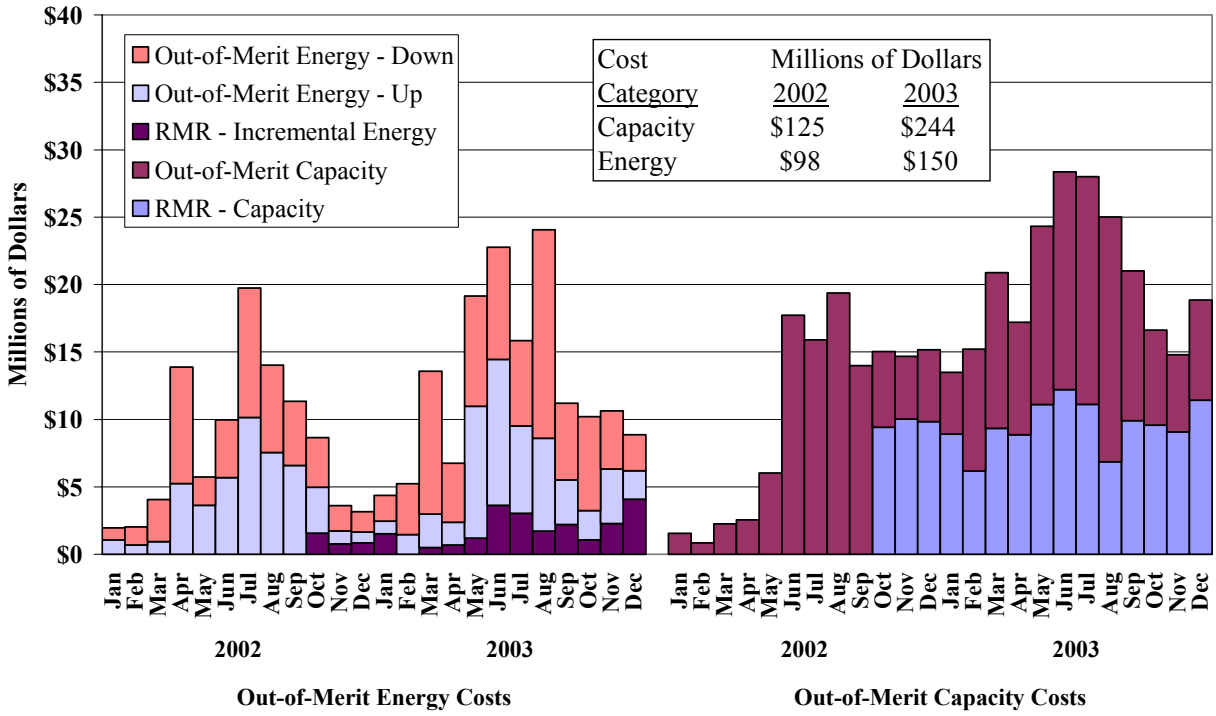
The pricing of the congestion rights is also important because the revenue from the auction of the congestion rights is primary means for the loads to receive the value of the transmission system that they pay for through regulated payments to transmission owners. In a perfectly efficient system with no uncertainty, the average congestion cost in real-time should equal the auction price of the congestion rights. In the real world, however, we would expect only reasonably close convergence with some fluctuations from year to year due to uncertainties.

In 2002, the annual auction for the congestion rights resulted in prices that substantially over-valued the congestion rights on the South to North and South to Houston interfaces. In 2003, the congestion rights auction prices for all of the interfaces decreased considerably, resulting in a much closer convergence with the actual value of the congestion rights. This likely indicates that, with the benefit of one year of market experience, participants have improved in their ability to forecast interzonal congestion and to value the congestion rights.

3. Local Congestion and Local Capacity Requirements

ERCOT manages local (intrazonal) congestion using out-of-merit dispatch ("OOME up" and "OOME down"), which causes units to depart from their scheduled quantities. When not enough capacity is committed to meet local reliability requirements, ERCOT sends OOMC instructions for offline units to start up to provide energy and reserves in the local area. In October 2002, RMR agreements were signed with certain generators needed for local reliability. When these units are called out-of-merit order, they receive revenues specified in the agreements rather than standard OOME or OOMC payments. Understanding the causes and patterns of local congestion is important. The following figure shows the out-of-merit energy and capacity costs, including RMR costs, for each month in 2002 and 2003.

Expenses for Out-of-Merit Capacity and Energy
2002-2003



In 2003, the costs associated with local congestion were almost $400 million as compared to costs of interzonal congestion of less than $25 million. This figure shows that OOME costs rose more than 50 percent from 2002 to 2003, while OOMC costs increased 95 percent. These costs include the costs for local balancing energy dispatched by SPD and reliability must run energy

and capacity costs. The largest source of these out-of-merit costs was local congestion in the North Zone. These local congestion costs in the North consisted of OOMC and OOME up costs in the import-constrained Dallas-Fort Worth area and OOME down costs in the export-constrained Northeast area. The latter has been addressed through the recent creation of a Northeast Zone, which should substantially reduce OOME down costs in 2004.

Some of the increase in the costs for managing local congestion in the West and South Zones is attributed to the introduction of RMR agreements in these areas that generally result in higher payments than the standard OOME and OOMC payments.

In the long-run, the adoption of the Texas Nodal markets would establish efficient prices that reflect all congestion within ERCOT and improve the allocation of the associated congestion costs. It would also improve participants' ability to hedge these costs since they would no longer be invisible in the market prices and recovered through uplift charges as is currently the case for local congestion. In the short-run, however, the report recommends that ERCOT consider creating a new zone for Dallas-Fort Worth as described above. This would likely cause much of the remaining OOME costs to be reflected in the balancing energy prices and would reduce uplift costs. However, we recognize that there are a number of important issues that would need to be considered in making this change in the short-run.

D. Balancing Energy Offers and Schedules

QSEs play an important role in the current ERCOT markets. QSEs must submit balanced schedules with scheduled resources that match their scheduled load. With the introduction of "relaxed balanced scheduling" in November 2002, there is no longer a requirement that the balanced schedules closely follow the QSE's actual load. The energy schedules are a primary input to determine the net supply and demand for balancing energy. In general, energy schedules that are less than the actual load result in balancing energy purchases while energy schedules higher than actual load result in balancing energy sales. QSEs also submit balancing energy offers to increase or decrease their energy output from the scheduled level. The balancing up offers correspond to the unscheduled output from the QSE's online and quick-start resources.

In addition to the forward schedules and offers, QSEs submit resource plans that provide a non-binding indication of the generating resources that the QSE will have online and producing

energy to satisfy its energy schedule and ancillary services obligations. The report evaluates the effects on the balancing energy market of the QSE's schedules, offers, and resource plans.

1. Scheduling Patterns

We evaluate forward scheduling patterns by comparing load schedules to actual real-time load. In the aggregate, load schedules tend to be under-scheduled by an average of 2 percent and by higher amounts under peak demand conditions. In some hours, the load is under-scheduled by 10 to 20 percent, which creates a sizable demand for balancing energy. This under-scheduling together with the balancing energy offer patterns described below sometimes result in large balancing energy price increases.

The North Zone is under-scheduled most significantly with the under-scheduling amounts ranging from 5 percent in low-load periods to 10 percent in high-load periods. A large portion of the under-scheduled load in the North Zone is satisfied by the same QSE's resources offered in the balancing energy market. We have not identified a reason for this scheduling pattern.

2. Hourly Schedules Changes

One of the most significant issues affecting the ERCOT balancing energy market is the changes in energy schedules that occur from hour to hour, particularly in hours when loads are changing rapidly (i.e., "ramping") in the morning and evening. The report shows that:

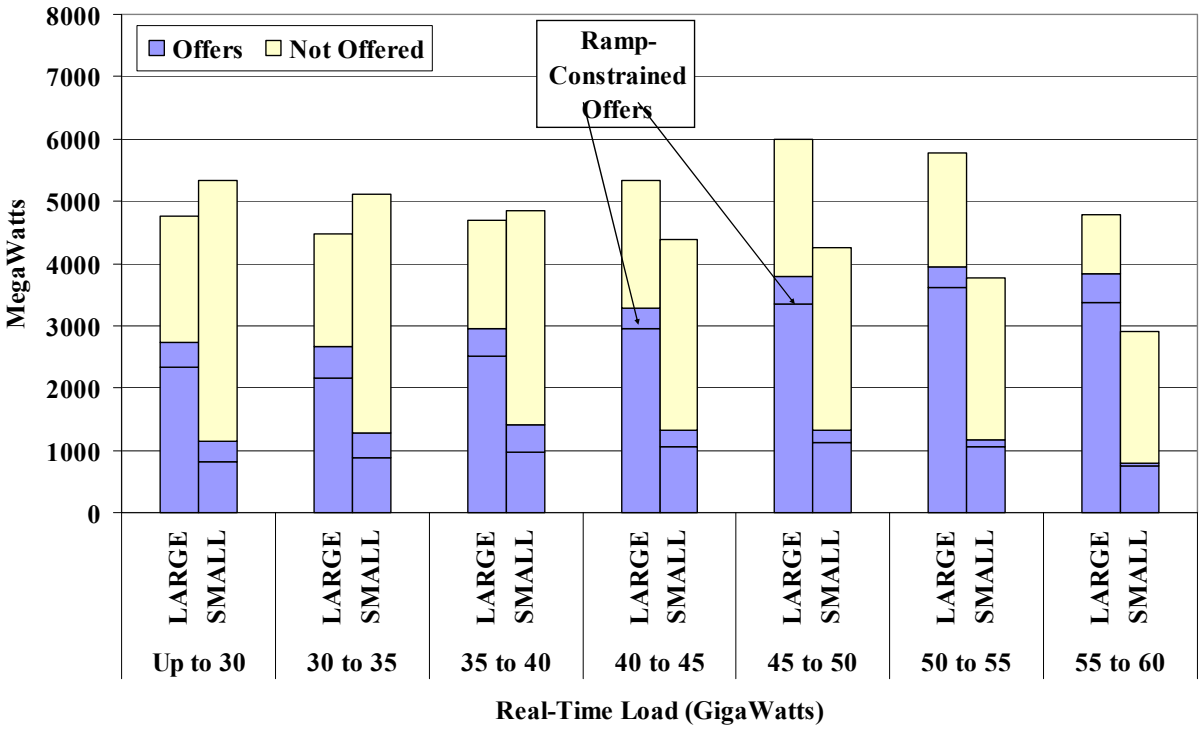
- In these ramping hours, the loads are generally moving approximately 300 to 400 MW each 15-minute interval.
- Although QSE's can modify their schedules each interval, most only change their schedules hourly, resulting in schedule changes averaging 1000 to 3000 MW in these hours (and sometimes significantly larger).
- The inconsistency between the changes in schedules and actual load in these hours places an enormous burden on the balancing energy market, resulting in the erratic pricing patterns shown above.
- The largest two QSEs schedule much more flexibly than the other QSEs and generally help to mitigate these problems.

To address this issue and improve the performance of the balancing energy market, the report recommends changes that may increase the willingness of QSEs to submit flexible schedules (i.e., schedules that change every 15 minutes).

3. Portfolio Offers in the Balancing Energy Market

The report evaluates the portfolio offers submitted by QSEs in the balancing energy market, including both the quantity and ramp rate of the offers (the amount of the offer that can be deployed in any single 15-minute interval). The figure below shows the total available energy versus the amount offered in the balancing energy market at different load levels by large QSEs and small QSEs.

Available Balancing Capacity Relative to Offers in 2003
Large versus Small Participants -- Daily Peak Load Hours



This figure and the other analysis of the portfolio offers indicate that:

- In general, less than half of the available capability is offered in the balancing energy market, although this amount rises to close to 75 percent at the very highest load conditions.
- The largest QSEs offer a much higher share of their available energy than smaller suppliers.
- The share of the available energy offered by the largest QSEs rises as load rises, which is not true of the other QSEs.
- Participants generally offer little more than the amount that can be deployed in a single interval (the additional amount is labeled "ramp-constrained offers" in the figure).

- However, the ramp limitations in the portfolio offers are much lower than the true physical ramp limitations of the individual generating units, which is a result of the portfolio bidding structure.
- This reduces the ability of the market to fully utilize the generating resources and can result in inefficient transitory fluctuations in balancing energy prices.

It is a significant concern that not all available capacity is offered into the market. Part of this problem can be attributed to the fact that gas turbine capacity is difficult to effectively offer in the balancing energy market, of which ERCOT currently has more than 3000 MW. The report includes a number of recommendations to address the portfolio ramp limitations and allow gas turbine capacity to be included in the portfolio offers, including:

- Considering modifications to the timing of the balancing energy market; and modification of portfolio offer validation procedures to ensure that they do not remove offline gas turbine capacity from the portfolio offers;
- Encouraging QSEs to submit multiple "sub-QSE" portfolio offers to reduce the ramp limitation effects of having all of a QSE's supply subject to a single ramp constraint.

4. Analysis of Resource Plans

The submission and use of resource plans raise general concerns because resource plans are not financially binding, yet they are used by ERCOT to determine when potentially costly out-of-merit commitment or dispatch actions must be taken. The report analyzes QSEs' resource plans to evaluate whether the market protocols may provide incentives for strategic conduct harmful to the market. In particular, the report evaluates units that are frequently committed out-of-merit or frequently dispatched out-of-merit. Such units receive additional payments from ERCOT and participants may engage in strategies to increase these payments. The analyses of these issues show:

- Units that are frequently committed out-of-merit and receive OOMC payments (two-thirds of which are located in Dallas-Fort Worth) are regularly indicated to not be online in the day-ahead resource plan, but were subsequently voluntarily committed by the QSE prior to real time.
- Units that frequently receive OOME up payments are generally scheduled significantly lower in the real-time resource plan than they actually operate in periods when they receive no OOME up instruction.

- Some units that frequently receive OOME down payments are scheduled higher in the real-time resource plan than they actually operate in periods when they receive no OOME down instruction.

These results indicate that the current market rules are providing incentives for QSEs to provide misleading resource plan information to increase the frequency and/or magnitude of the OOME and OOMC payments. To address the concerns raised by these results, the report recommends two alternatives to the current markets that would mitigate these issues:

- Consider the creation of a zone for Dallas/Ft. Worth to allow a large share of the local congestion to be priced more efficiently and transparently. This would also provide superior economic signals to guide investment in generation and transmission in that area. We understand that the effects on existing bilateral contracts, the need for measures to mitigate market power in the area, and the equity implications of such a change would all need to be carefully evaluated.
- If the foregoing is infeasible, the local congestion costs should be directly assigned to the load in the constrained area. Although it is not as transparent or efficient, it would significantly improve the incentives of QSEs with generation and load in the constrained areas.

In the long-term, the most comprehensive solution for all of these issues would be to implement nodal energy markets. Nodal markets are much less vulnerable to these types of incentive concerns because they send efficient price signals to generators that reflect all network constraints, virtually eliminating the need for out-of-merit dispatch actions and associated payments.

E. Analysis of Competitive Performance

The report evaluates two aspects of market power, structural indicators of market power and behavioral indicators that would signal attempts to exercise market power. The structural analysis in this report focuses on identifying circumstances when a supplier is "pivotal", i.e., when its generation is needed to serve the ERCOT load and satisfy the ancillary services requirements. This analysis leads to the following results and conclusions.

- When load obligations are considered, the suppliers in ERCOT are rarely pivotal.
- However, because a large portion of the available energy from online resources is routinely not offered in the balancing energy market, we found that a supplier was pivotal in 10 percent of all hours and in 24 percent of hours when real-time load exceeded 40 GW.

- Although the balancing market may not reflect traditional market power, the factors described above that prevent full utilization of the available energy in the balancing market make it more vulnerable to manipulation.

While structural market power indicators are very useful in identifying potential market power issues, they do not address the actual conduct of market participants. Accordingly, we analyze physical and economic withholding in order to further evaluate competitive performance of the ERCOT market. Based on the analyses conducted in this area, the report finds little evidence of systematic physical or economic withholding of generating resources during 2003. However, it is important to recognize that these analyses of physical and economic withholding only evaluate trends or patterns that would raise broad competitive concerns. Isolated instances of significant physical or economic withholding must be identified on a case-specific basis.

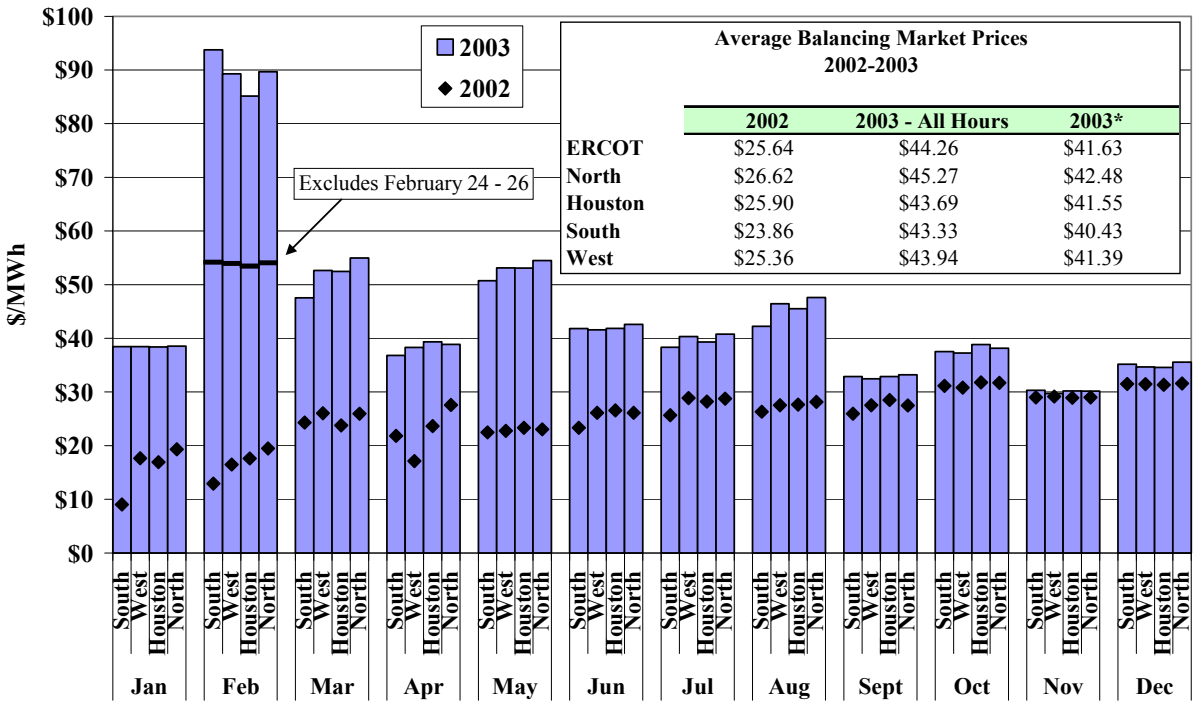
I. REVIEW OF MARKET OUTCOMES

A. Balancing Energy Market

1. Prices During 2003

Balancing energy prices in 2003, particularly in the first eight months of the year, were substantially higher than in 2002. As the analysis in this report will show, this is primarily due to considerable increases in natural gas prices. To summarize the overall price levels during the past two years, Figure 1 shows the load-weighted prices from the balancing energy market for each ERCOT zone in 2003 and 2002.

Figure 1: Average Balancing Energy Market Prices 2002 & 2003



Average Balancing Market Prices 2002-2003

	2002	2003 - All Hours	2003*
ERCOT	$25.64	$44.26	$41.63
North	$26.62	$45.27	$42.48
Houston	$25.90	$43.69	$41.55
South	$23.86	$43.33	$40.43
West	$25.36	$43.94	$41.39

*Excludes February 24-26.

This figure shows that the load-weighted average prices throughout ERCOT were more than 70 percent higher in 2003 than in 2002.[2] The effects of tight conditions in the natural gas market were most significant on February 24-26 when balancing energy prices exceeded $900 per

[2] The load-weighted average prices are calculated by weighting the balancing energy price in each interval and zone by the total zonal loads in that interval. This is not consistent with prices reported elsewhere that are weighted by the balancing energy procured in the interval.

MWh. Figure 1 also shows that electricity prices in February were 66 percent higher than they would have been without the three days of extreme prices. Further, these three days increased the average prices for the year by 6.3 percent as shown in the table in Figure 1.

The fact that such a small number of high-priced hours can have a significant effect on the average prices over the entire year illustrates the significant influence that price spikes generally have on the economic signals provided by the market. It also reinforces the importance of ensuring that price spikes occur efficiently – i.e., that prices rise efficiently during periods of legitimate shortages and that price spikes do not result from withholding in the absence of a shortage.

Figure 1 also shows that the price differences between the zones tend to be relatively small, reflecting only moderate amounts of interzonal congestion. In both years, the North Zone exhibited the highest average prices and the South Zone exhibited the lowest. The average difference in prices between these two zones was approximately 12.0 percent in 2002 and 4.5 percent in 2003.

The balancing energy market is a spot energy market through which a very small share of the power produced in ERCOT is transacted, which is typical for a spot market. Although most power is purchased through bilateral forward contracts, outcomes in the balancing energy market are very important because of the expected pricing relationship between spot and forward markets. Unless there are barriers that prevent arbitrage of the prices in the spot and forward markets, the prices in the forward market should be directly related to the prices in the spot market (i.e., the spot prices and forward prices should converge over the long-run).[3] Hence, artificially-low prices in the balancing energy market will translate to artificially-low forward prices and, likewise, price spikes in the balancing energy market will increase prices in the forward markets.

The next analysis evaluates the total cost of serving load in the ERCOT market. In addition to the costs of energy, loads incur costs associated with operating reserves, regulation, and uplift.

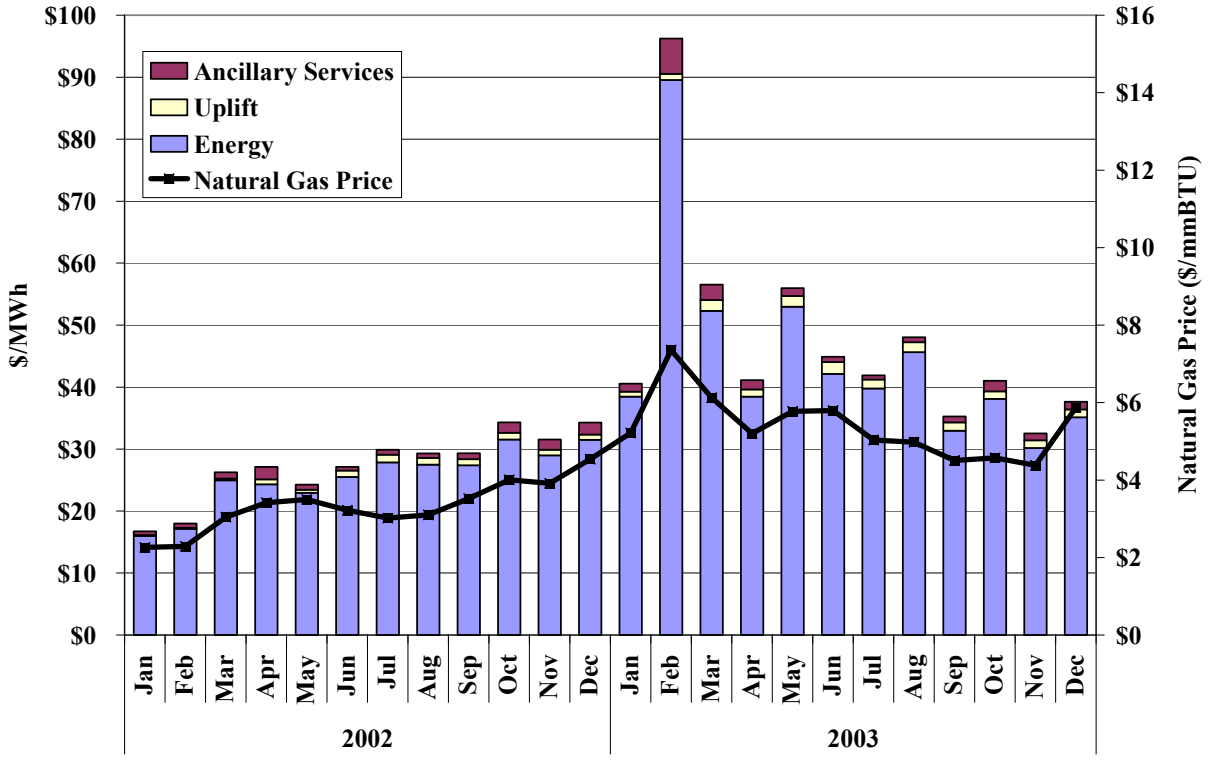
[3] See Hull, John C. 1993. *Options, Futures, and other Derivative Securities*, second edition. Englewood New Jersey: Prentice Hall, p. 70-72.

(As discussed more below, uplift costs are associated with payments by ERCOT for out-of-merit dispatch and out-of-merit commitment.) We have calculated an average all-in price of electricity for ERCOT that is intended to reflect energy costs as well as these additional costs. The monthly average of this metric for all of ERCOT in 2002 and 2003 is shown in Figure 2.

The components of the all-in price of electricity include:

- Energy costs: Balancing energy prices are used to estimate energy costs, under the assumption that the price of bilateral energy purchases converges with balancing energy prices over the long-term, as discussed above.
- Ancillary services costs: These are estimated based on the demand and prices in the ERCOT markets for regulation, responsive reserves, and non-spinning reserves.
- Uplift costs: Uplift costs are assigned market-wide on a load-ratio share basis.

Figure 2: Average All-in Price for Electricity in ERCOT 2002 to 2003



With the exception of February 2003, all-in prices are fairly stable from month-to-month. What is notable, however, is the relatively weak relationship of energy and reserves prices to load

levels. This relationship and the reasons for it are examined in more detail below. The anomalous all-in prices in February were the result of electricity price spikes over three days (February 24-26) when prices rose as high as $990 per MWh in the balancing energy market. These price spikes occurred in response to a spike in natural gas prices and unusually high loads associated with a period of extremely cold weather.

The Market Oversight Division ("MOD") of the Public Utility Commission of Texas ("PUCT") conducted a thorough review and analysis of this event, issuing a report containing its findings in May 2003.[4] In its report, the MOD concludes that the primary factors contributing to the price spikes on February 24-26 were:

- The inability of generators to procure sufficient natural gas on February 24;
- The failure of many Qualifying Scheduling Entities ("QSEs") to provide ERCOT with timely updates on the status of specific generating units; and
- "Hockey stick" bidding by one QSE.

We agree with these conclusions regarding the extreme weather event that occurred on February 24-26 that resulted in price spikes during this period.

In general, however, we find that the amount of available energy that is not offered in the ERCOT balancing energy market is generally a more significant issue than "hockey-stick" bidding. Whether unusually high-priced offers are not competitively justified (i.e., hockey-stick offers) or do represent a competitive offer (i.e., marginal costs) for a specific output range[5], the lack of full participation in the balancing energy market can cause these high-priced offers to be deployed during periods when no shortage exists.

Figure 2 also shows that natural gas prices were a primary driver of the trends in electricity prices in 2002 and 2003. This is not surprising given that natural gas is the predominant fuel in ERCOT, especially among the generating units that most frequently set the balancing energy prices. Natural gas prices increased in 2003 by more than 65 percent from 2002 levels on

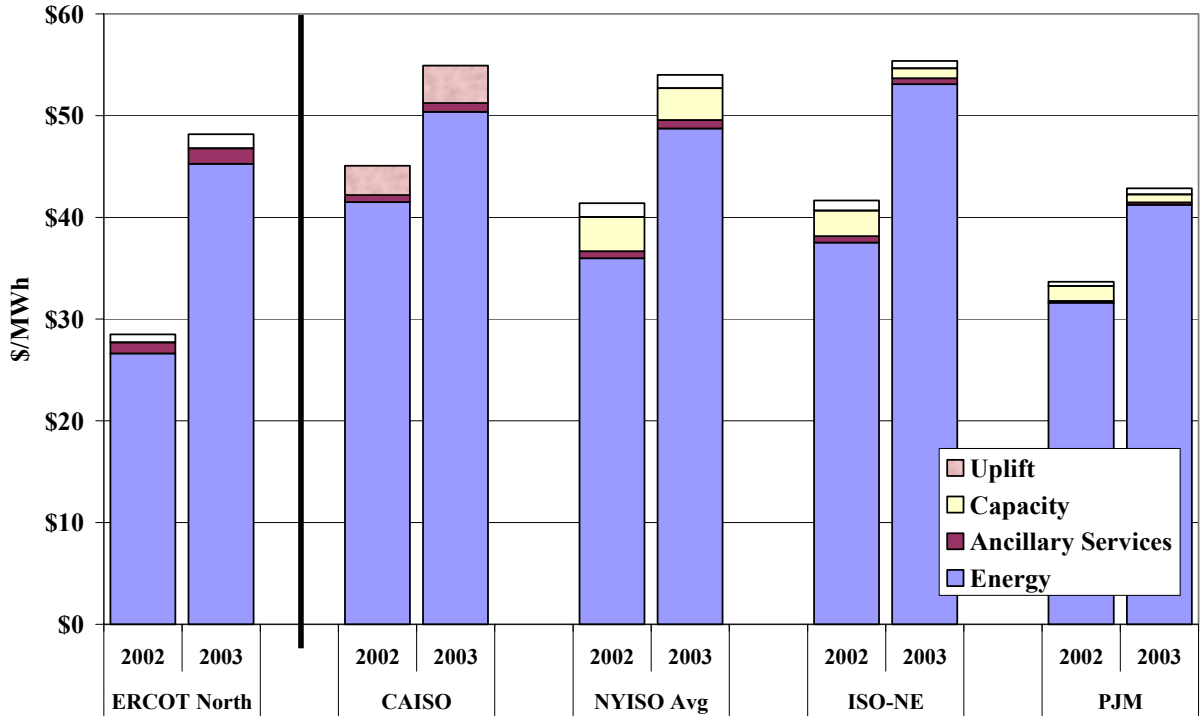
4 Public Utility Commission of Texas, Market Oversight Division, "Market and Reliability Issues Related to the Extreme Weather Event on February 24-26, 2003," report filed in Project Number 25937 (May 19, 2003).

5 For example, steam units can experience extremely high incremental costs and increased risks when operating at emergency levels above their normal output range.

average while the all-in price for electricity increased by 72 percent. Although higher natural gas prices were among the most important reasons the all-in energy prices increased in 2003, there were other important factors. For example, Figure 2 shows that uplift costs generally associated with local congestion increased in 2003. This is analyzed and discussed in detail later in the report. The report also examines scheduling and balancing energy offer patterns that may have contributed to the higher all-in prices in 2003.

To provide some perspective on the outcomes in the ERCOT market, our next analysis compares the all-in price metrics for ERCOT and other electricity markets. Figure 3 compares the all-in prices for the five major centralized wholesale markets in the U.S.: (a) ERCOT, (b) California ISO, (c) New York ISO, (d) ISO New England, and (e) PJM ISO. For each region, the figure reports the average cost (per MWh of load) for (a) energy, (b) ancillary services (reserves and regulation), (c) capacity markets (if applicable), and (d) uplift for economically out-of-merit resources.

Figure 3: Comparison of All-In Prices across Markets 2002 - 2003



Each market experienced a substantial increase in energy prices from 2002 to 2003 due to increased fuel costs, particularly natural gas costs. Although the markets vary substantially in the portion of their generating capacity that is fired by natural gas, these units are on the margin and set the wholesale spot prices in the majority of hours for all markets shown.

The largest increase in all-in prices between 2002 and 2003 occurred in ERCOT. In 2002, ERCOT exhibited the lowest all-in price -- 18 percent lower than the next lowest-priced market. In 2003, the all-in price in PJM, which experienced the lowest increase in prices from 2002 to 2003, was lower than in ERCOT. Natural gas-fired generation is on the margin less frequently in PJM than any of the other markets because PJM has access to large quantities of coal-fired generation within PJM itself and in the Midwest. In spite of a substantial rise in 2003, the all-in prices in the ERCOT region remain relatively low due in part to its substantial resource margin.

Figure 3 also shows that ERCOT had higher costs for reserves and regulation than other markets in 2003 stemming primarily from higher prices for these services, which is examined in the next subsection. Uplift costs were also substantial in ERCOT in 2003, second only to uplift costs in the CAISO market. This is expected because both ERCOT and CAISO operate zonal electricity markets in which intrazonal congestion costs are recovered through uplift charges. NYISO, ISO-New England, and PJM operate locational marginal pricing ("LMP") markets where all congestion costs are included in the locational prices and collected through the energy market.

The all-in energy prices in ERCOT do not include capacity market costs that exist in all of the other markets (CAISO reports capacity costs together with reserves and regulation costs) which makes the comparison difficult. Although capacity markets are not a necessary component of a wholesale electricity market, they do reduce the market's reliance on high energy prices that reflect shortage conditions as a mechanism to provide the primary economic signal to govern new investment and retirement decisions. Additionally, a capacity market will allow a region to sustain a higher resource margin than would exist without it. Capacity market costs ranged from $0.79 to $3.15 per MWh of load in 2003 in the Northeast markets. Capacity costs are highest in New York due to the locational capacity requirements in New York City, which is currently close to being deficient in capacity.

Our final analysis of all-in prices (shown in Figure 4) indicates how the market costs vary by ERCOT zone.

Figure 4: Average All-In Price of Electricity by Zone 2002 to 2003

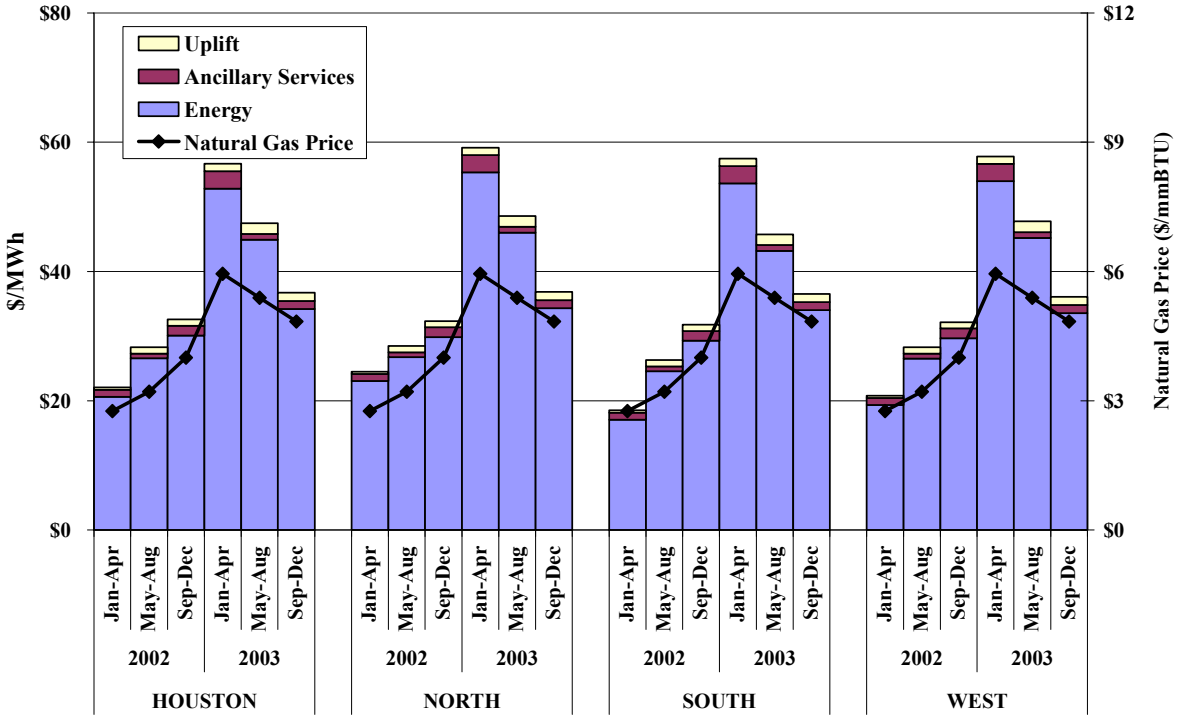
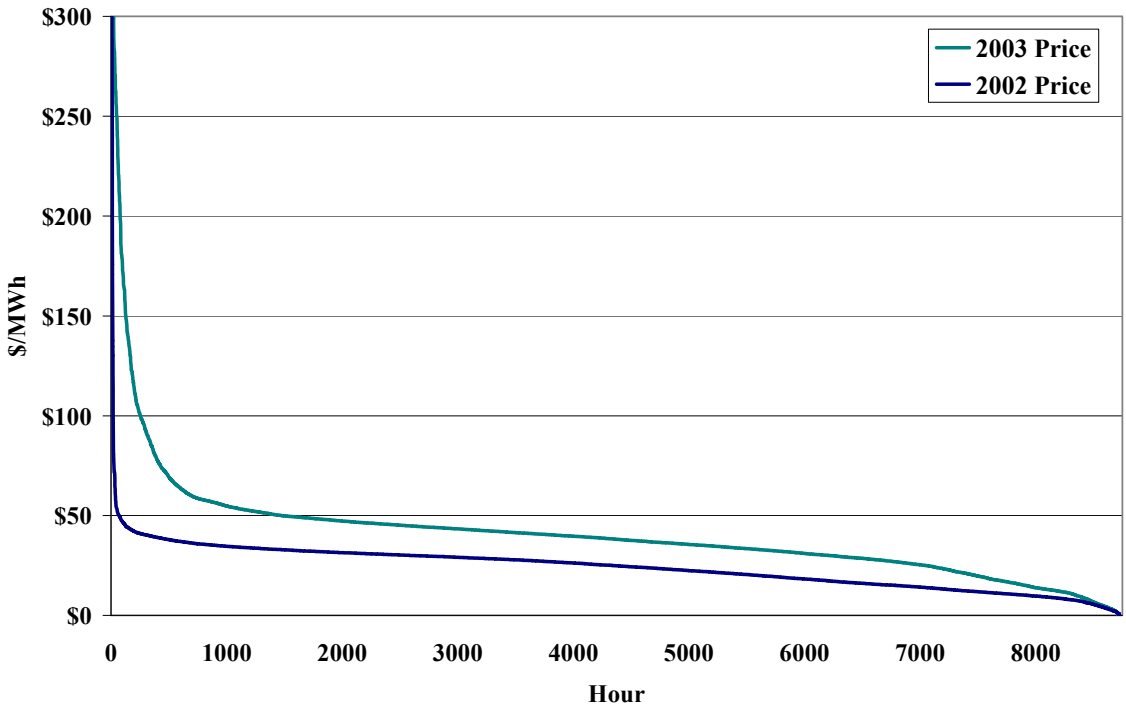


Figure 4 shows that there is relatively little difference in prices between zones. The largest interzonal price differences occurred at the beginning of 2002. The figure also shows that the uplift costs were comparable in size to reserves and regulation costs over 2002 and 2003. Uplift costs were higher than reserves and regulation costs during the summer months and usually lower during other periods.

The analyses above show the average balancing energy prices by month and year. The next analysis shows the hourly balancing energy prices. Figure 5 presents price duration curves for the balancing energy market in 2002 and 2003. A price duration curve indicates the number of hours that the price is at or above a certain level. The prices in this figure are hourly load-weighted average prices for the ERCOT balancing energy market.

Figure 5: ERCOT Price Duration Curve
2002 & 2003



This figure shows that prices were relatively low in 2002, exceeding $50 in only 73 hours. In contrast, almost 1,500 hours (one-third of the peak hours) exhibited prices higher than $50 in 2003. This clearly illustrates the effect of higher fuel prices, which increased electricity prices over nearly the entire range of hours for the year. This occurs because higher natural gas prices raise the marginal production costs of the generating units that set the prices in most hours. In other words, natural gas price increases will cause suppliers to increase their balancing energy offers, resulting in higher balancing energy prices in most hours. Figure 5 also shows that prices were substantially higher in the highest-priced hours. The increases in these hours are not fully explained by increases in natural gas prices.

To better observe the highest-priced hours, Figure 6 shows a narrower set of data that focuses on the highest-priced five percent of hours. The prices in these hours play a significant role in providing economic signals to invest in new and retain existing generation.

Figure 6: Price Duration Curve
Top Five Percent of Hours – 2002 & 2003

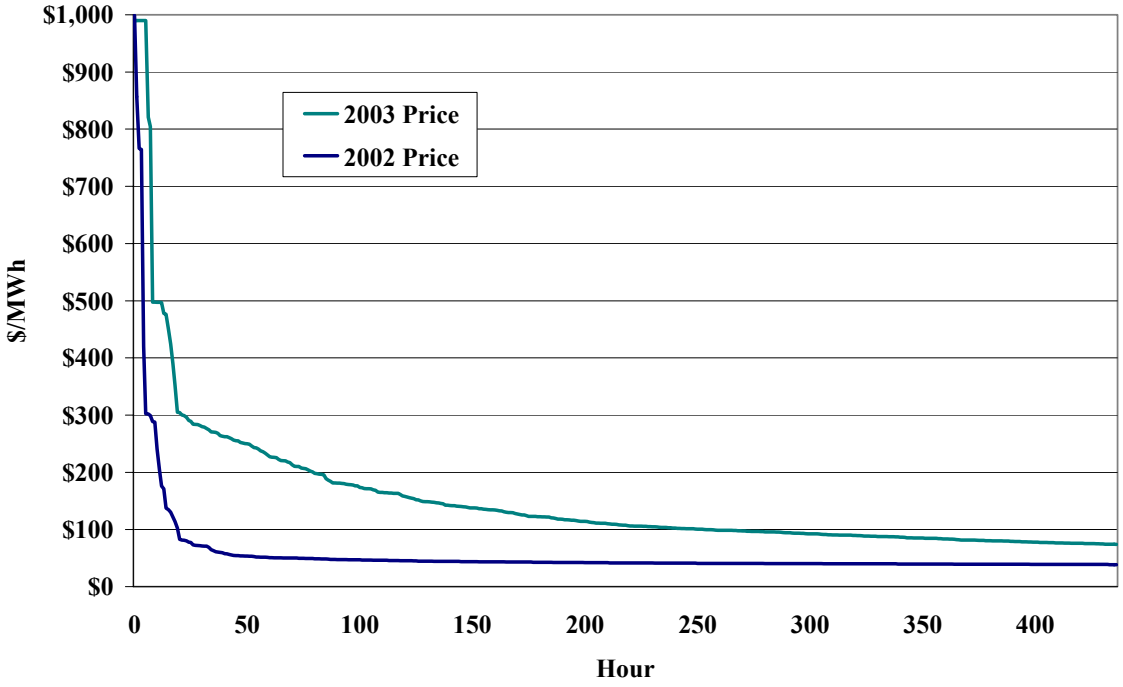


Figure 6 shows a more dramatic difference between 2002 and 2003 in the highest-priced hours than in all other hours. In 2002, there were only 20 hours with prices over $100 per MWh and only 12 hours with prices over $200 per MWh. In contrast, prices in 2003 exceeded $100 per MWh in 254 hours and exceeded $200 per MWh in 80 hours. Normally, one would expect the highest-priced hours to occur during the summer peak-demand conditions. However, high prices occurred as often during the spring as during the summer in 2003. Although some of the high prices during the spring were due to the winter peak in natural gas prices that extended through the spring, the high prices in the spring frequently occurred under moderate load conditions and could not be explained by natural gas prices. The fact that these prices occurred despite a substantial excess of generating capacity in ERCOT raises issues regarding the efficiency and competitiveness of the balancing energy market that is examined below.

2. Price Convergence

One indicator of market performance is the extent to which forward and real-time prices converge over time. In ERCOT, there is no centralized day-ahead market so prices are formed in the day-ahead bilateral contract market. The real-time spot prices are formed in the balancing

energy market. Forward prices will converge with real-time prices when two main conditions are in place: a) there are low barriers to shifting purchases and sales between the forward and real-time markets; and b) sufficient information is available to market participants to allow them to develop accurate expectations of future real-time prices. When these conditions are met, market participants can be expected to arbitrage predictable differences between forward prices and real-time spot prices by increasing net purchases in the lower-priced market and increasing net sales in the higher-priced market. This will tend to improve the convergence of the forward and real-time prices.

We believe these two conditions are largely satisfied in the current ERCOT market. One important step taken to address the first condition (i.e., to reduce barriers between the markets), was the implementation of relaxed balanced schedules in November 2002. By allowing QSEs to increase and decrease their purchases in the balancing energy market, they should be better able to arbitrage forward and real-time energy prices. While this should result in better price convergence, it should also reduce QSEs' total energy costs by allowing them to increase their energy purchases in the lower-priced market.[6]

It should be noted, however, that the current balancing energy market does not reveal the full value of energy in the ERCOT market. Intrazonal constraints associated with "local congestion" are not reflected in balancing energy prices, which tends to undervalue energy in locally-constrained areas. Instead, these congestion costs are borne in market-wide uplift charges that cannot be hedged through forward energy contracts. Hence, neither the balancing energy prices nor the forward energy prices will include the costs of managing local congestion.

There are several ways to measure the degree of price convergence between forward and real-time markets. In our analysis, we measure two aspects of convergence. The first method investigates whether there are systematic differences in prices between forward markets and the real-time market. The second tests whether there is a large spread between real-time and forward prices on a daily basis.

---

6 The volatility in balancing energy prices, which increased in 2003, creates risk that may cause some participants to be willing to pay a premium to purchase energy in the bilateral markets and should, therefore, result in a premium in the bilateral market prices above the balancing energy prices over time.

To determine whether there are systematic differences between forward and real-time prices, we evaluate the difference between the average prices in the two markets in 2003. This measures whether persistent and predictable differences exist between forward and real-time prices, which participants should arbitrage over the long-term.

In order to measure the short-term deviations between real-time and forward prices, we also calculate the average of the absolute value of the difference between the forward and real-time prices on a daily basis during peak hours. This indicates the volatility of the daily price differences, which may be large even if the average difference over the year is low.[7] These two statistics are shown in Table 1 for both 2002 and 2003.

Table 1: Convergence Between Forward and Real-Time Energy Prices 2002 to 2003

	Average price of power on weekdays from 6 AM to 10 PM			
	Day-Ahead Price	Balancing Energy Price	Avg Price Difference	Absolute Avg Difference
2002	$29.06	$26.57	$2.49	$4.95
2003	$46.56	$48.96	-$2.41	$12.66

Note: Day-Ahead Price based on Megawatt Daily peak day-ahead prices when five or more trades were reported.

The table indicates that there was a sizable premium in the day-ahead prices in 2002 compared to the balancing energy prices. In 2003, prices were generally higher and there was a closer correspondence between day-ahead and real-time prices in percentage terms. This is likely due, in part, to the introduction of relaxed balanced schedules that increased participants' flexibility to arbitrage prices between the day-ahead forward market and the balancing energy market as discussed above. In addition, the reversal in the relationship between day-ahead and real-time prices was largely caused by the price spikes in late February that were unexpected and increased

[7] For instance, if forward prices are $70 per MWh on two consecutive days while real-time prices are $40 per MWh and $100 per MWh on the two days, the absolute price difference between the forward market and the real-time market would be $30 per MWh on both days, while the difference in average prices would be $0 per MWh.

the average balancing energy prices by almost $3 per MWh. But for this event, the day-ahead and real-time prices would have been close to equal.

Although this measure of price convergence improved considerably in 2003, it is important to recognize that this does not indicate that the balancing energy prices were efficient. The performance of the balancing energy market, including the conduct of the market participants, is evaluated in subsequent sections of this report. In addition, we will be evaluating ERCOT's market operations and congestion management procedures in a subsequent report to be issued later this year.

Table 1 also shows that the average absolute price difference more than doubled in 2003. This is partially due to the higher prices that prevailed in 2003. As a percentage of the balancing energy price, the absolute price difference increased from 18 percent to 26 percent. Much of this increase is associated with the larger number of unforeseen price spikes that occurred in 2003 under relatively mild load conditions. In contrast, there were virtually no price spikes in 2002. There are a number of factors that have contributed to the increased volatility in the balancing energy market in 2003. These factors are identified and evaluated later in this report. The results in this section indicate that the effectiveness of the ERCOT market in achieving convergence between the day-ahead bilateral prices and the balancing energy prices improved from 2002, although the volatility of the balancing energy market must be further examined.

3. Volume of Energy Traded in the Balancing Energy Market

In addition to its role in providing a vital signal of the value of power for market participants entering into forward contracts, the balancing energy market plays a role in governing real-time dispatch. This section examines the volume of activity in the balancing energy market.

The amount of energy traded in ERCOT's balancing energy market is small relative to overall energy consumption. Most energy is purchased and sold through forward contracts that insulate participants from volatile spot prices. Because forward contracting does not precisely match generation with real-time load, there will be residual amounts of energy bought and sold in the balancing energy market. Moreover, the balancing energy market enables market participants to make efficient changes from their forward positions, such as replacing relatively expensive generation with lower-priced energy from the balancing market.

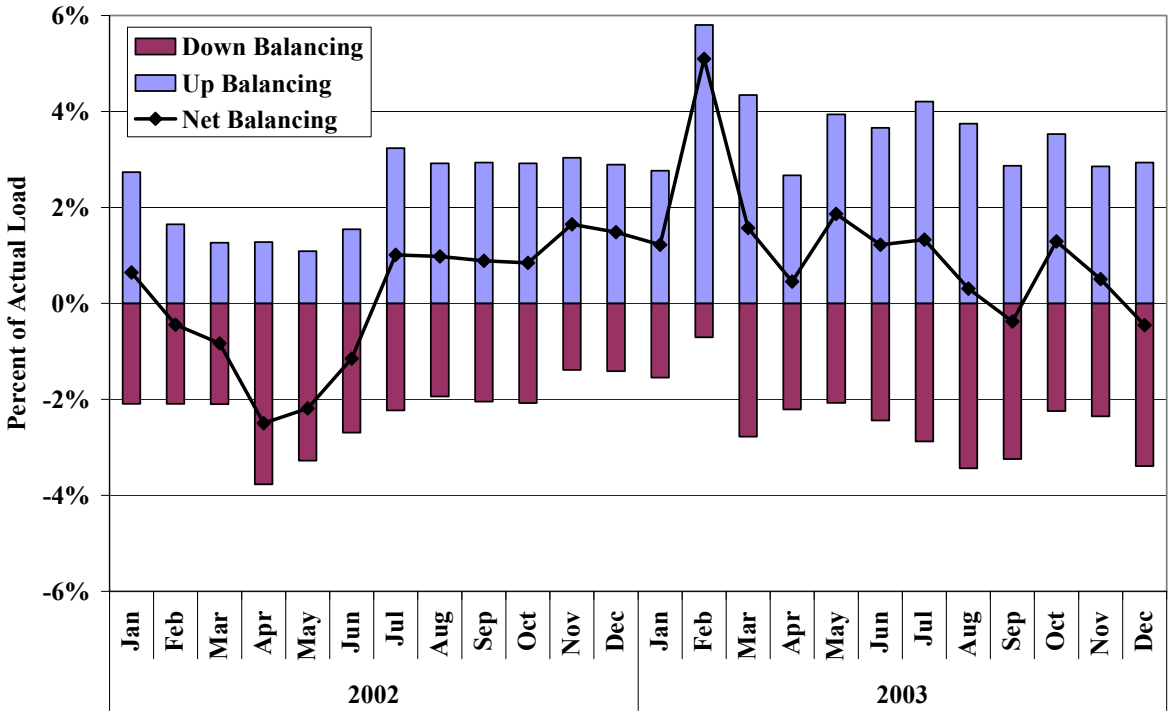
Hence, the balancing energy market will improve the economic efficiency of the dispatch of generation to the extent that market participants make their resources available in the balancing market. In the limit, if all available resources were offered competitively in the balancing energy market (to balance up or down), the prices in the current market would be identical to clearing all power through a centralized spot market (even though most of the commodity currently settles bilaterally). It is rational for suppliers to offer resources in the balancing energy market even when they are fully contracted bilaterally because they can increase their profit by reducing their output and supporting the bilateral sale with balancing energy purchases. Hence, the balancing energy market should govern the output of all resources, even though only a small portion of the energy is settled through the balancing market.

In addition to their role in governing real-time dispatch, balancing energy prices also provide a vital signal of the value of power for market participants entering into forward contracts. As discussed above, the spot prices emerging from the balancing energy market should directly affect forward contract prices assuming that the market conditions and market rules allow the two markets to converge efficiently.

This section summarizes the volume of activity in the balancing energy market. Figure 7 shows the average quantities of balancing up and balancing down energy sold by suppliers in each month, along with the net purchases or sales (i.e., balancing up energy minus balancing down energy).

Figure 7 shows that the total volume of balancing up and balancing down energy as a share of actual load increased by 33 percent, averaging 4.5 percent of actual load in 2002 and 6.0 percent in 2003. The average in 2003 remains close to 6 percent even when February 2003 (when the extreme weather event occurred) is excluded. Thus, there was a general increase in trading through the balancing energy market in 2003.

Figure 7: Average Quantities Cleared in the Balancing Energy Market 2002 & 2003



One reason for the increase in balancing energy trading may be the implementation of relaxed balanced schedules in November 2002. This change allows market participants to intentionally schedule more or less than their anticipated load, and to buy or sell in the balancing energy market to satisfy their actual load obligations. This has allowed the balancing energy market to increasingly operate like a centralized energy spot market and has contributed to improved price convergence, although the increase in the volume of energy traded through the balancing energy market is not as large as some expected.

Figure 7 shows that February 2003 is the only month when net balancing up sales exceeded five percent of total sales. From February 24th to March 6th, suppliers sold an average of 2,465 MW of net balancing up energy, approximately 7.8 percent of load. This high level of purchases was due in part to natural gas curtailments and generation outages that compelled some of the load-serving entities to turn to the balancing energy market to purchase additional energy to serve load. These factors were identified by MOD in a report that focused on the most extreme portion of this period, from February 24 to 26.[8] This report also showed, however, that there was

[8] Public Utility Commission of Texas, Market Oversight Division, *Op cit.*

available energy not offered in the balancing energy market, which is common in ERCOT and is an issue we investigate below.

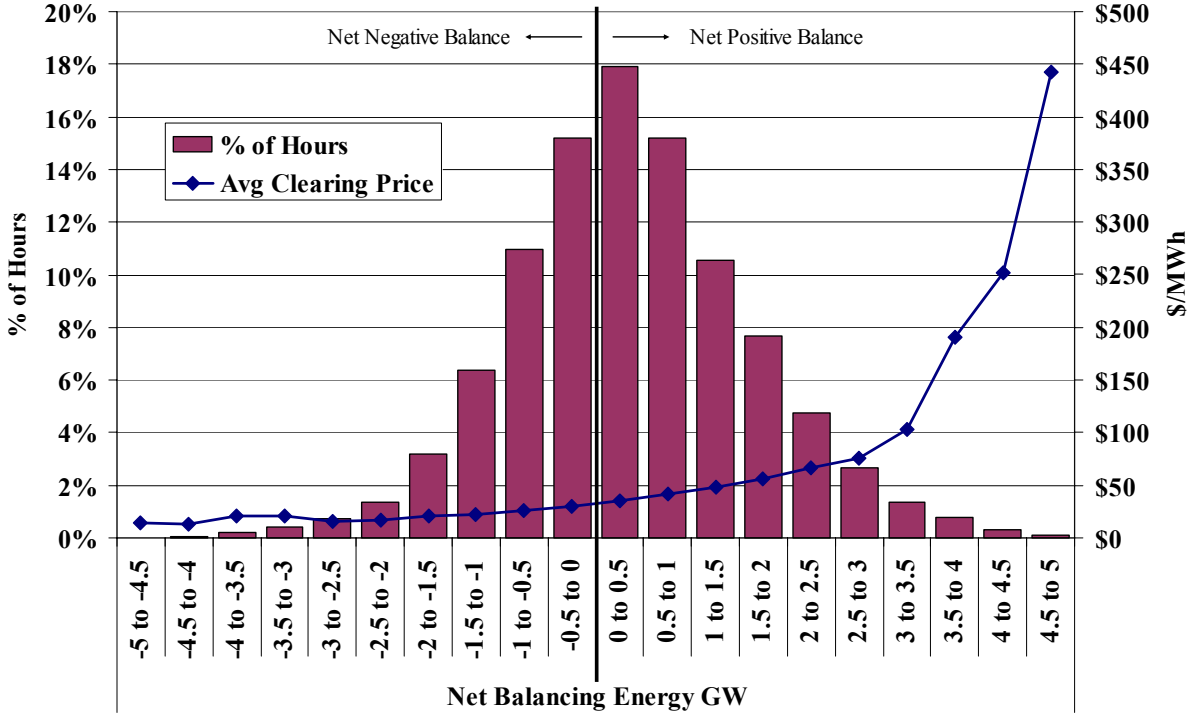
There are two aspects of the increase in trading volume shown in Figure 7 that should be considered separately. First, large quantities of balancing up and balancing down energy that are deployed simultaneously to clear "overlapping" balancing offers will necessarily result in efficiency improvements. This activity reflects the displacement of higher-cost resources with lower-cost resources, lowering the overall costs of serving load and allowing the balancing energy price to more accurately reflect the marginal value of energy.

The second aspect of the increase in trading volume that is important is the increase in *net* balancing energy. When large quantities of net balancing-up or net balancing-down energy are scheduled, it indicates that QSEs are systematically under-scheduling or over-scheduling load relative to real-time needs. One reason this can occur is to arbitrage the forward energy and balancing energy markets. Figure 7 shows that the average monthly net balancing energy volume has fluctuated significantly over the two years, although it has been positive in all but two months since July of 2002. This is consistent with an increase in market participant purchases in the balancing energy market in response to the low balancing energy prices relative to the daily forward prices that prevailed in 2002.

However, if large hourly under-scheduling or over-scheduling occurs suddenly, the balancing energy market can lack the ramping capability and sometimes the volume of energy offers necessary to achieve an efficient outcome. In these cases, large net balancing energy purchases can lead to transient price spikes when excess capacity exists but is not available in the 15-minute time frame of the balancing energy market. The remainder of this sub-section and the next sub-section will examine in detail the patterns of over-scheduling and under-scheduling that have occurred in the ERCOT market and the effects that these scheduling patterns have had on balancing energy prices.

To provide a better indication of the frequency with which net purchases and sales of varying quantities are made from the balancing energy market, Figure 8 presents a distribution of the hourly net balancing energy.

Figure 8: Magnitude of Net Balancing Energy and Corresponding Price 2003



Each bar in Figure 8 shows the portion of the hours during 2003 when balancing energy purchases or sales were in the range shown on the x-axis. For example, the figure shows that the quantity of net balancing energy traded was between zero and positive 0.5 gigawatts (i.e., loads were under-scheduled and net purchasing of balancing energy occurred) in 18 percent of the hours in 2003.

Figure 8 shows that the mean of the distribution is positive with a symmetrical distribution of net balancing energy purchases that is skewed towards net balancing up purchases. This is consistent with Figure 7 which showed that there were net balancing up quantities on a monthly average basis in 10 of the 12 months in 2003. Figure 8 also shows that almost 60 percent of the hourly observations show net purchases or sales between -1.0 gigawatts and 1.0 gigawatts.[9] Hence, there were many hours when the net balancing energy traded was relatively low, indicating that in many hours the total scheduled energy is close to the actual load.

[9] One gigawatt corresponds to roughly 3 percent of the average actual load in ERCOT.

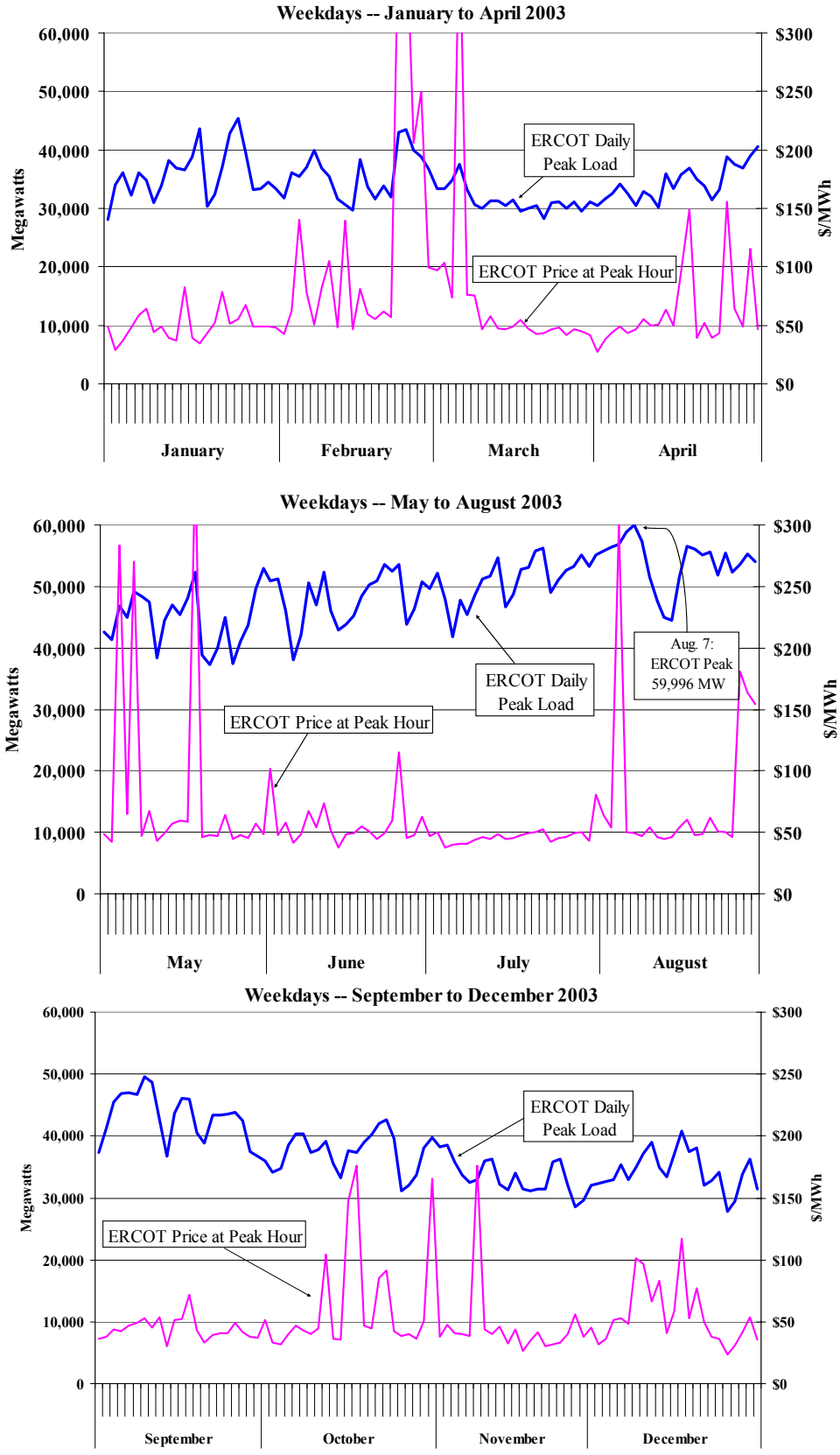
Figure 8 also shows the average balancing energy prices corresponding to each level of balancing energy volumes. In an efficiently functioning spot market, there should be little relationship between the balancing energy prices and the net purchases or sales. Instead, one should expect that prices would be primarily determined by more fundamental factors, such as actual load levels and fuel prices. However, this figure indicates a relatively clear relationship, showing the balancing energy prices increasing as net balancing energy volume increases. This provides a preliminary indication that the balancing energy market is relatively thinly traded, which can undermine its efficiency. We analyze the potential reasons for this apparent relationship in the next sub-section.

4. Determinants of Balancing Energy Prices

The prior section shows that the level of net sales in the balancing energy market appears to play a significant role in explaining the balancing energy prices. In this section, we examine this relationship in more detail, as well as the role of more fundamental determinants of balancing energy prices, such as the ERCOT load and fuel prices.

Figure 9 shows the average balancing energy price and the actual load in the peak-load hour of each weekday during 2003. The figure shows that the highest prices (e.g., greater than $100/MWh) do not reliably correspond to the highest load levels. Indeed, the price was close to $50 per MWh at the system peak, lower than in many other lower-load days. Likewise, prices throughout the summer were generally not positively correlated to peaks in load.

Figure 9: Daily Peak Loads and Prices



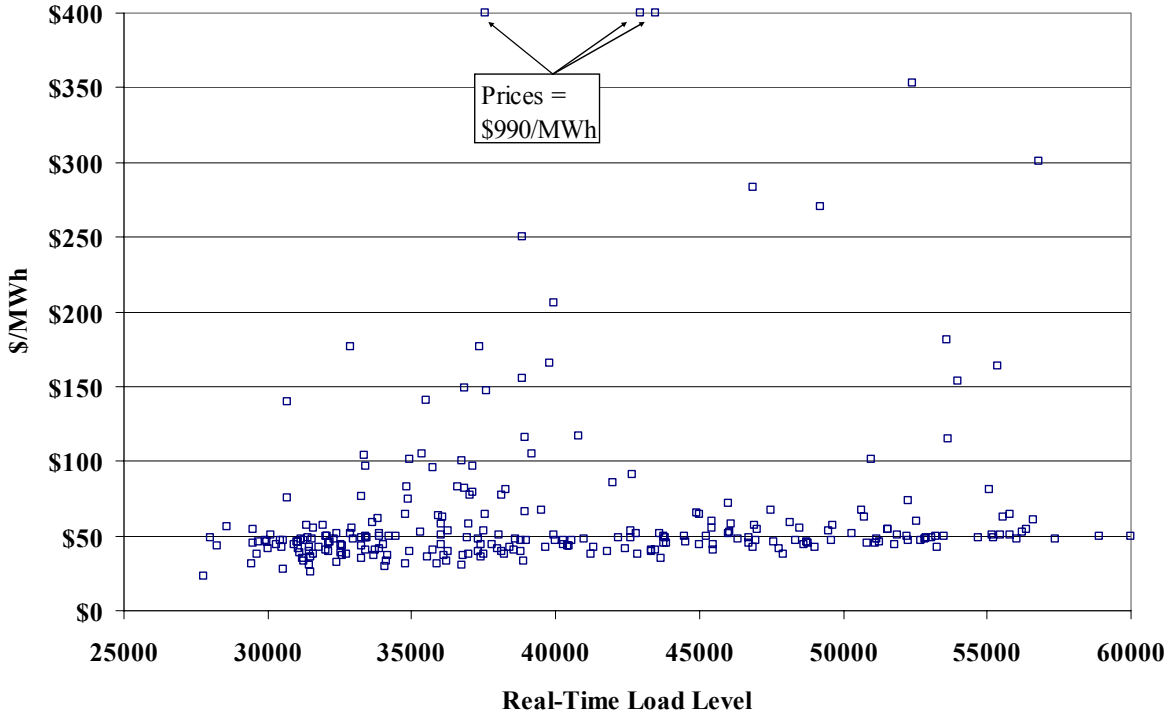
The highest prices during the year occurred in late February, a period which was associated with unusually cold weather and very tight conditions in the natural gas markets. Although the February loads were not close to the annual peak level, they were higher than in any other period in the winter. In addition to contributing to the higher electricity loads, the cold weather during this period affected the availability of a number of generating units and the availability of natural gas. As discussed previously, the detailed analysis of this event by the Market Oversight Division identifies these factors, as well as the submission of "hockey-stick" offers by one QSE as the primary causes of these price spikes.[10]

Excluding the price spikes of late February and early March associated with tight conditions in the natural gas market, the highest prices occurred in early and mid-May. These increases correspond to peak load days during the month of May. Although the load was close to 10 GW lower during these days than at the annual peak in August, they occurred during a time when a substantial amount of generation was scheduled to be out-of-service for maintenance. Hence, the actual load did play a significant role in contributing to these price increases, although the overall relationship of prices to actual load remains erratic.

To further examine the relationship of the actual load in ERCOT to the balancing energy prices, Figure 10 shows the same data as Figure 9, but plots the average balancing energy prices versus the daily peak loads in ERCOT irrespective of time.

[10] Public Utility Commission of Texas, Market Oversight Division, *Op Cit*.

Figure 10: ERCOT Balancing Energy Price vs. Real-Time Load
Weekdays -- Peak Load Hour -- 2003



These prices are generally tightly clustered around $50 per MWh. If one examines the relatively high prices, i.e., those greater than $100 per MWh, there is little discernable relationship between these occurrences and the actual load in ERCOT. In fact, most of these prices occur when load is less than 40 GW. Although short-term peaks in actual load can sometimes help explain relatively high balancing energy prices, this relationship is not as reliable as we expected. Alternatively, the analysis above indicates that the net volumes of energy purchased in the balancing energy market appear to be an unexpectedly important determinant of the balancing energy prices.

To more precisely examine these relationships, we performed some econometric tests to identify the extent to which net sales in the balancing energy market determine the balancing energy prices. We control for other fundamental factors that influence the balancing energy price, including the natural gas price and the adequacy of supply in real-time. The adequacy of supply is captured in the quantity of excess real-time capability, which is equal to the amount of in-service generating capacity (i.e., not on outage) minus the quantity of real-time demand.

We estimated the following equation for all intervals between March 15 and December 31, 2003:[11]

$$\text{Balancing Energy Price} = a + b_1*\text{Excess Capability} + b_2*\text{Net Balancing Energy} + b_3*\text{Natural Gas Price}$$

This analysis will indicate whether any or all of these factors are a unique determinant of the balancing energy price, controlling for the other factors. Table 2 shows the results of the estimation.

Table 2: Linear Regression of Balancing Energy Price
All Intervals – March 15 to December 31, 2003

Dependent Variable = Balancing Energy Price ($/MW)	Total R-squared = 0.70 R-squared Excluding AR Term = 0.15	
Independent Variables	**Parameter Estimate**	**Standard Error**
Constant	45.60	3.64
Excess Capability in Real-Time (GW)	-0.99	0.05
Net Upward Balancing Energy (GW)	10.40	0.18
Natural Gas Price ($/Mmbtu)	3.09	0.61

Note: Prais-Winsten autoregression method used to correct for serial correlation. All parameter estimates are statistically significant at the 99% confidence interval or higher.

These results show that each of the three independent variables is statistically significant, which confirms our observations from the analyses above. The estimated parameter for Net Balancing-up Energy shows that prices are relatively responsive to changes in the net balancing energy purchases, increasing by more than $10 per MWh for each GW increase in net purchases. As discussed above, this raises concerns regarding the efficiency of the balancing energy market because a statistically significant relationship between the balancing energy prices and the net balancing energy purchases would not be expected in a well-functioning market.

The estimated parameter for Natural Gas Price indicates a predictable relationship between natural gas prices and electricity prices. The coefficient implies that the electricity price would be expected to increase approximately $3 per MWh due to a $1 per MMBTU increase in the

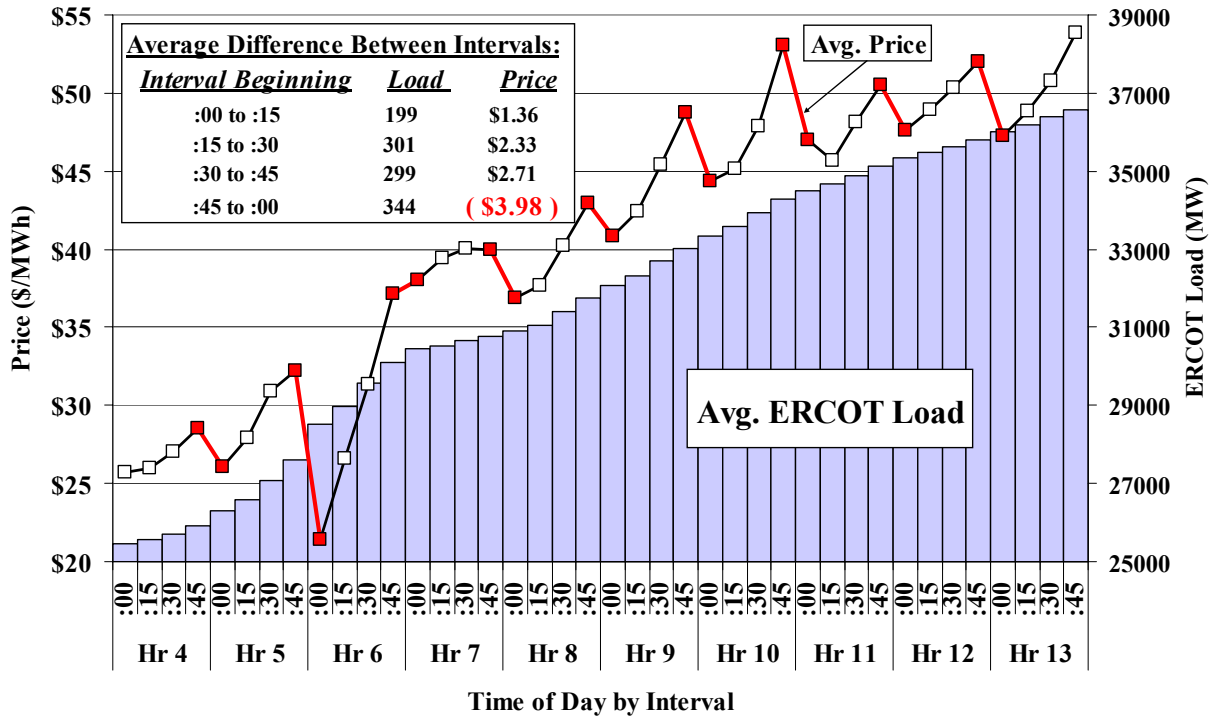
[11] The sample excludes the period before March 15th in order to eliminate the natural gas crisis. Data on unit availability during this period is not reliable. Including this period could bias the parameter estimates.

natural gas price. Although the statistical significance of the natural gas prices is expected due to the reliance on natural gas-fired units in ERCOT, the magnitude of the effect is less than half what we expected. Because gas-fired units are on the margin in the Texas market (i.e., setting balancing energy prices) in a large share of the hours, we would expect that the change in electricity prices would reflect the change in the marginal costs of the natural gas units, determined primarily by their heat rates. The estimated parameter in our regression implies a heat rate of 3,090 BTU/KWh, while the incremental heat rates of most natural gas units would range from 7,000 to 12,000. This indicates that the balancing energy prices are less responsive to changes in natural gas prices than we would expect in a market where natural gas-fired units are the marginal source of generation.

Finally, the regression results show the expected negative relationship between the quantity of excess capability in real-time and balancing energy prices. For each reduction in excess capability of one gigawatt, the model estimates approximately a one dollar per MWh increase in balancing energy prices. In other words, when less excess capability is available in real-time, balancing energy prices tend to be higher as one would expect. As load levels increase, the quantity of excess capability decreases by an equivalent amount. This relatively small coefficient on the excess capability variable confirms the observations from the figures that there is a weak relationship between load levels and the balancing energy prices.

To further examine how the prices relate to actual load levels, the final analysis in this subsection shows the average balancing energy prices by interval during the hours each day when load is increasing or decreasing rapidly (i.e., when load is ramping up and ramping down). ERCOT load rises during the day from an average of approximately 25 GW to 37 GW. This usually occurs over a ten-hour period. Thus, the change in load averages 1,200 MW per hour (300 MW per interval) during the morning and early afternoon. Figure 11 shows the average load and balancing energy price in each interval from 4 AM through 1 PM in 2003. The price is plotted as a line in the figure while the average load is shown with vertical bars.

Figure 11: Average Clearing Price and Load by Time of Day
Ramping-Up Hours – 2003



The figure shows that the load steadily increases in every interval and prices generally move upward from about $26 per MWh at 4:00 AM to $54 per MWh at 1:45 PM. If actual load were the primary determinant of energy prices, the balancing energy prices would rise gradually as the actual load rises. However, Figure 11 shows a distinct pattern in the balancing energy prices over the intervals. The balancing energy price rises throughout each hour and drops substantially in the first interval of the next hour. In the figure, particular emphasis is shown on the transition from one hour to the next hour. The average price change from the last interval of one hour to the first interval of the next hour is -$3.98 per MWh.

A similar pattern is observed at the end of the day when load is decreasing. ERCOT load decreases in the evening more quickly than it increases early in the day. Most of the decrease occurs over a six hour period, averaging a decrease of 1,800 MW per hour (450 MW per interval) during the late evening. Figure 12 shows this decrease in load by interval, together with the average balancing energy prices for the intervals from 9 PM to 3 AM.

Figure 12: Average Clearing Price and Load by Time of Day
Ramping-Down Hours – 2003

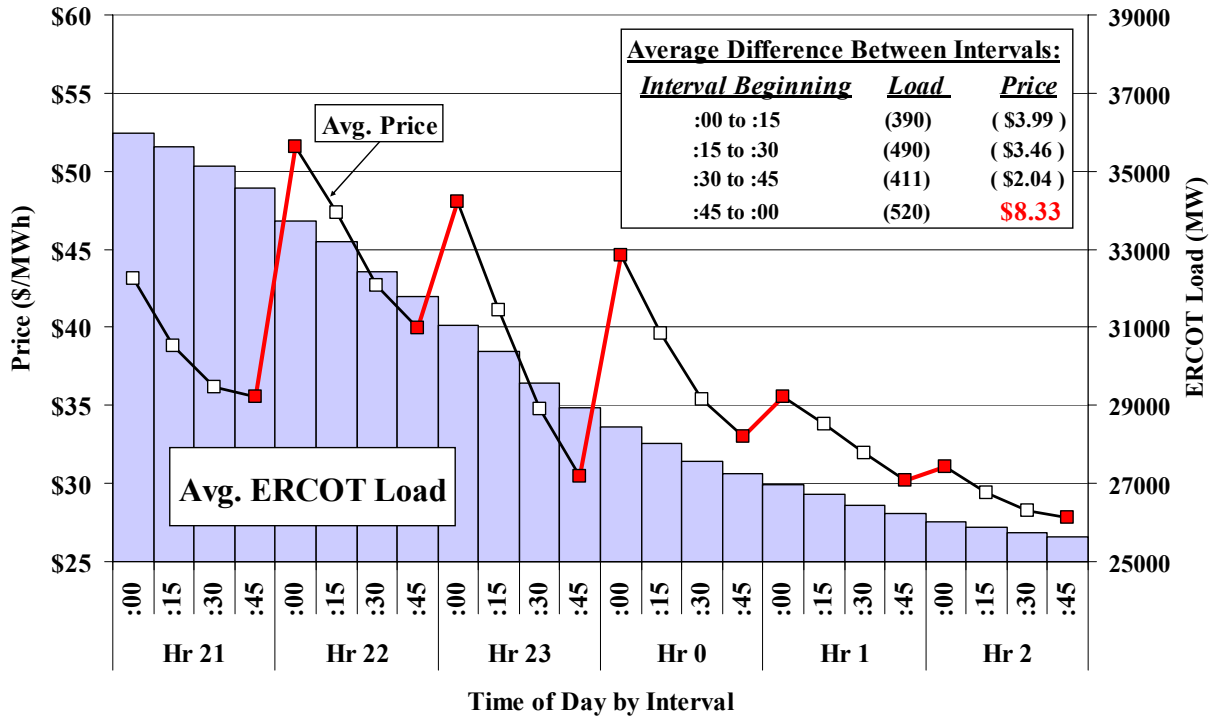


Figure 12 shows that although the balancing energy prices decrease over these intervals, they follow a similar pattern as exhibited in the ramping-up hours. The balancing energy price decreases in each interval of the hour before rising substantially in the first interval of the following hour. The balancing energy price increases by an average of $8.33 per MWh from the last interval of one hour to the first interval of the next hour during this period.

These figures show that this pattern of balancing energy prices by interval is not explained by changes in actual load. Rather, changes in balancing energy deployments by interval underlie this pricing pattern. Sizable changes in balancing energy deployments occur between intervals, particularly in the first interval of the hour. These changes are associated with large hourly changes in energy schedules. These scheduling and pricing patterns are examined in detail in Section II below.

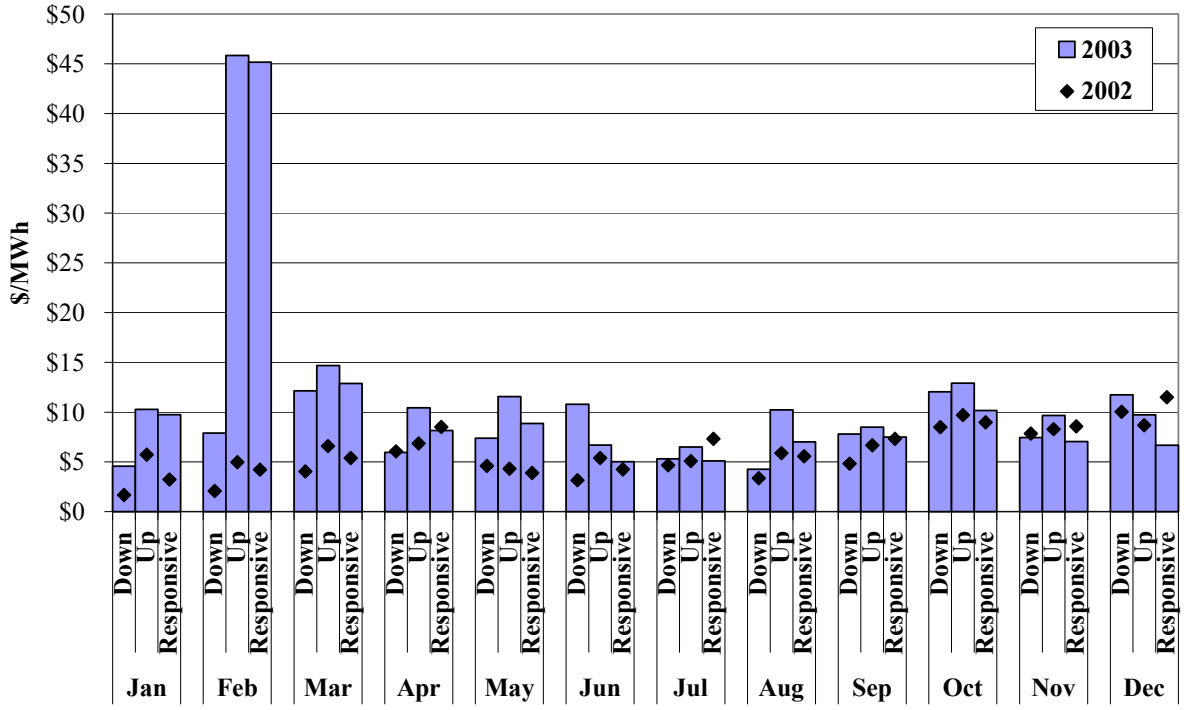
B. Ancillary Services Market Results

The ancillary services markets are generally cleared day-ahead. The primary ancillary services are up regulation, down regulation, and responsive reserves. ERCOT may also procure non-spinning reserves as needed. QSEs may self-schedule ancillary services or purchase their required ancillary services through the ERCOT markets. This section reviews the results of the reserves and regulation markets in 2003.

1. Reserves and Regulation Prices

Our first analysis in this section provides a summary of the ancillary services prices over the past two years. Figure 13 compares the monthly ancillary services prices between 2002 and 2003. Average prices for each ancillary service are weighted by the quantities required in each hour.

Figure 13: Monthly Average Ancillary Service Prices 2002 & 2003



This figure shows that ancillary services prices were generally higher in 2003, particularly during the first half of the year. Much of this increase can be attributed to the increase in energy prices that occurred over the same timeframe. Because these markets are conducted prior to the balancing energy market, participants must include their expected costs of foregone sales in the

balancing energy market in their offers for responsive reserves and regulation. Both providers of responsive reserves and up regulation can incur such opportunity costs if they reduce the output from economic units to make the capability available to provide these services.

Likewise, providers of down regulation can incur opportunity costs in real-time if they receive instructions to reduce their output, although these expected costs are likely to be lower than the costs of providing up regulation. This is consistent with the prices in 2003 when up regulation prices exceeded down regulation prices in all months except two, as shown below in Figure 13.

The figure also shows that the prices for up regulation exceeded prices for responsive reserves in all months in 2003. This is consistent with expectations because a supplier must incur the same opportunity costs to provide both services, while providing up regulation can generate additional costs. These additional costs include (a) the costs of frequently changing output, and (b) the risk of having to produce output when regulating at balancing energy prices that are less than the unit's variable production costs. This pricing relationship between responsive reserves and up regulation was not consistent in 2002 -- five of the twelve months in 2002 exhibited responsive reserves prices that were higher than the up regulation prices. This improvement in price consistency occurred despite the fact that regulation requirements were reduced in 2003, which would tend to reduce regulation prices.

The next analysis, shown in Table 3, compares the average prices for responsive reserves and non-spinning reserves over the past two years. This table shows the average prices (weighted by the quantity of non-spinning reserves required) for these two services in hours when ERCOT procured non-spinning reserves. It also shows average prices for 2002, with and without April 29 and 30 when prices ranged above $990 for 13 hours. We excluded these prices from the table because they tend to obscure the overall price relationship between the two services.

Non-spinning reserves were purchased in approximately 18 percent of the hours during 2002 and 25 percent of hours during 2003. Like the relationship between regulation and responsive reserves prices, responsive reserves prices should exceed non-spinning reserves prices because responsive reserves are a higher quality of reserves. Resources capable of providing responsive reserves can also be used to provide non-spinning reserves, but the reverse is not true. Hence, the price for non-spinning reserves should never exceed the price of responsive reserves.

Table 3: Responsive Reserves and Non-Spinning Reserves Prices 2002-2003

	2002*	2003
Non-Spinning Reserve Price	$6.30	$9.82
Responsive Reserve Price	$8.37	$10.73

* Excludes April 29-30, 2002. Including these days, the average prices were $14.43 for Non-spin and $9.19 for Responsive.

Table 3 shows the expected relationship between average prices for responsive and non-spinning reserves. However, non-spinning reserves prices were still higher than responsive reserves prices in 17 percent of hours during 2002 and 35 percent of hours during 2003. Although non-spinning reserves are a lower quality product than responsive reserves, these pricing results may not be as counterintuitive as they appear because providing non-spinning reserves may actually be more costly to provide for some resources than responsive reserves.

When a resource providing reserves is actually deployed to produce energy, the deploying QSE is paid for the output at the balancing energy price. There is no guarantee that the balancing energy price will be higher than the cost of dispatching the resource. When the balancing energy price is lower, no additional payment is made to the QSE. In fact, it is likely most units providing reserves have production costs higher than the balancing energy price because these units are the lowest cost providers of reserves (because these units incur no lost profits by not producing energy). Hence, these units will be running at a loss if they are deployed. Hence, the risk of losses associated with reserve deployments should be included in the operating reserves offer prices by suppliers. The two determinants of the expected value of these losses are: (a) the average difference between the resource's production cost and the balancing energy price, and (b) the probability of being deployed for energy. It is the second factor that can cause the marginal cost of supplying non-spinning reserves (and hence the clearing prices for non-spinning reserves) to be higher than for responsive reserves.

In 2003, less than 0.1 percent of the responsive reserves were actually deployed (not including under frequency relays) while 6.5 percent of non-spinning reserves were actually deployed. Therefore, the expected value of the deployment costs may cause the provision of non-spinning reserves to be more costly for some units than responsive reserves, which could contribute to counterintuitive results in some hours. In general, the purpose of operating reserves is to protect

the system against unforeseen contingencies (e.g., transmission line or generator outages), rather than for meeting load. The balancing energy market deployments in the 15-minute timeframe and regulation deployments in the 6-second timeframe are the primary means for meeting the load requirements.

However, in cases where the resources in the balancing and regulation markets may not be sufficient to satisfy the energy demand while meeting the responsive reserve requirement, we understand that ERCOT will frequently procure and deploy non-spinning reserves. This process is a means for ERCOT to implement supplemental generator commitments to increase the supply of energy in the balancing energy market. While supplemental generator commitments can be necessary for a variety of reasons, this is not a typical or desirable use of an operating reserve market. In our subsequent report reviewing the ERCOT market operations, we will examine this process and may provide recommendations to improve the supplemental commitment procedures.

Ultimately, the objective in the long-run should be to jointly-optimize each of the ancillary services markets with the energy market. In a market where ancillary services are jointly optimized with energy, the marginal cost of providing non-spinning reserves can never be higher than the marginal cost of providing responsive reserves. As in ERCOT, a jointly optimized market will deploy non-spinning reserves more frequently than responsive reserves because responsive reserves are more critical for reliability and are therefore more valuable. However, when non-spinning reserves are deployed in the context of a jointly optimized market, there is no risk that the clearing price will be insufficient for these units to recover their production costs since they will contribute to setting the energy prices. A jointly-optimized market recognizes the energy offer prices for all resources that are dispatched.

There is a pending project that will modify the procurement process and simultaneously clear the ancillary services markets (but not jointly optimize them with the balancing energy market), which should help ensure efficient pricing relationships between the services. This change is likely to result in increased prices in the responsive reserve market to reflect the higher marginal costs of providing non-spinning reserves. Since the costs of providing non-spinning reserves

may be partly attributable to the deployment procedures discussed above, it will be particularly important to consider potential improvements to these procedures.

Responsive reserves prices were relatively high during 2003, averaging almost $11 per MWh. Figure 14 shows how the annual average prices in ERCOT compare to the responsive reserve prices in the California, PJM, and New York wholesale markets.

Figure 14: Responsive Reserves Prices in Other RTO Markets 2002 and 2003

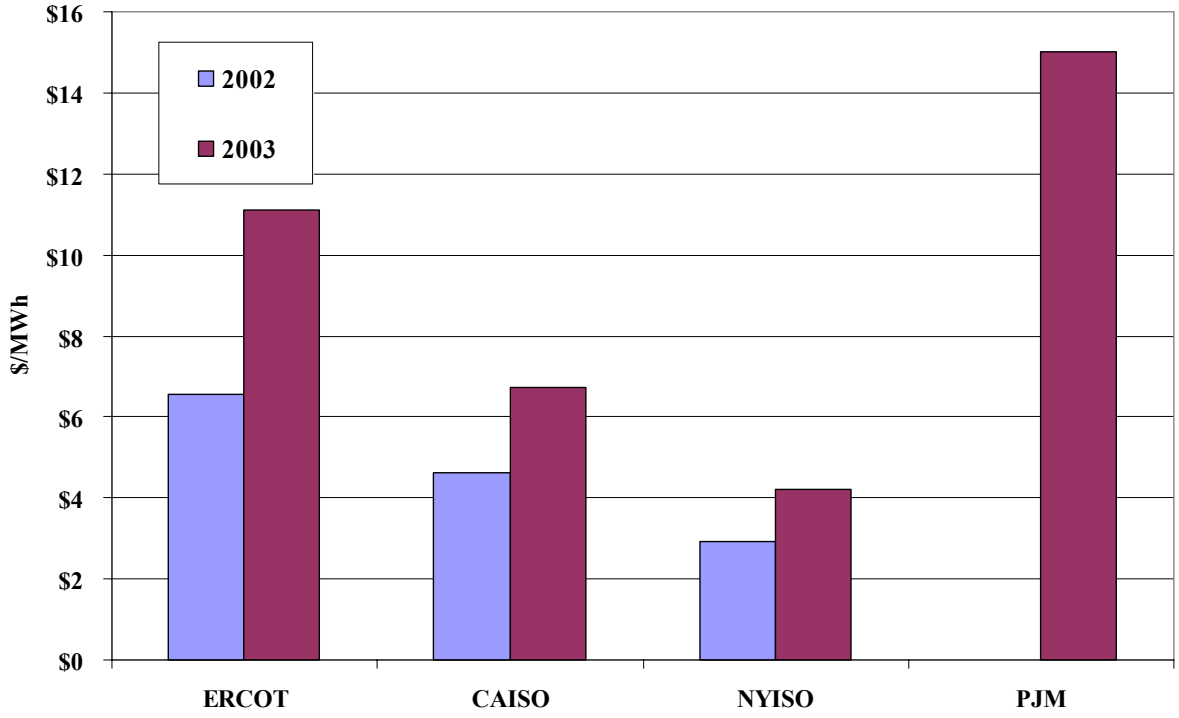


Figure 14 shows that the responsive reserve prices in ERCOT were substantially higher than comparable prices in California and New York, but lower than prices in PJM. Only 2003 prices are shown for PJM which instituted a market for spinning reserves in December 2002. There are a number of reasons why the responsive reserve prices in ERCOT are higher than prices in some of the other regions. First, ERCOT procures substantially more responsive reserves relative to its load than New York, which satisfies a large share of its operating reserve requirements with non-spinning reserves and 30-minute reserves rather than responsive reserves (i.e., 10-minute spinning reserves). However, more than one-third of ERCOT's responsive reserves are satisfied

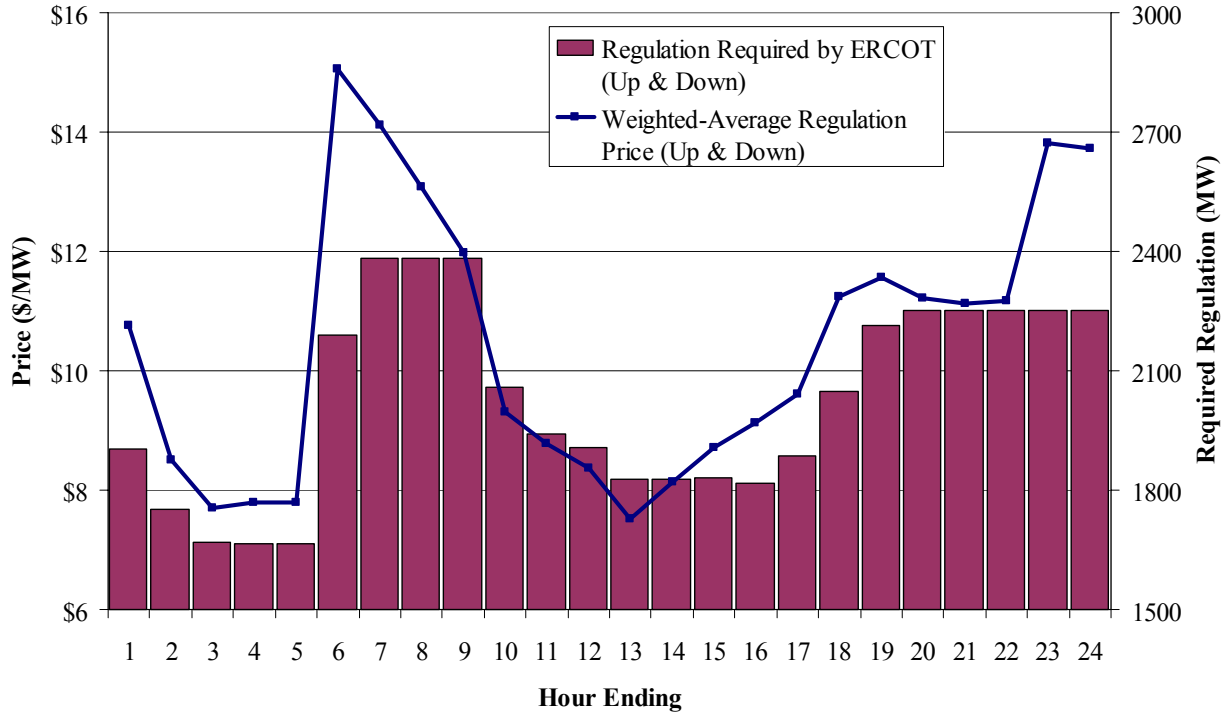
by demand-side resources offered at very low prices, which serves to offset the fact that ERCOT procures a higher quantity of responsive reserves.

Second, ERCOT does not jointly-optimize ancillary services and energy markets, like California and PJM. The lack of joint-optimization will generally lead to higher ancillary services prices because participants must incorporate in their offers the potential costs of pre-committing resources to provide reserves or regulation. These costs include the lost profits from the energy market when it would be more profitable to provide energy than ancillary services. Lastly, the offer patterns of market participants can influence these clearing prices. These offer patterns are examined in the next section.

Our next analysis evaluates the variations in regulation prices. The market dispatch model runs every fifteen minutes and produces instructions based on QSE-scheduled energy and balancing energy market offers, while regulation providers keep load and generation in balance by adjusting their output continuously. When load and generation fluctuate by larger amounts, more regulation is needed to keep the system in balance. This is particularly important in ERCOT due to the limited interconnections with adjacent areas, which results in much greater variations in frequency when generation does not precisely match load. Movements in load and generation are greatest when the system is ramping, thus ERCOT generally needs 25 percent more regulating capacity during ramping hours. When demand rises, higher cost resources must be employed and prices should increase.

Figure 15 shows the relationship between the quantities of regulation demanded by ERCOT and regulation price levels. This figure compares regulation prices to the average regulation quantity (both up and down regulation together) procured by the hour of the day. Regulation prices are an average of up and down regulation prices weighted by the quantities of each that are procured.

Figure 15: Regulation Prices and Requirements by Hour of Day 2003

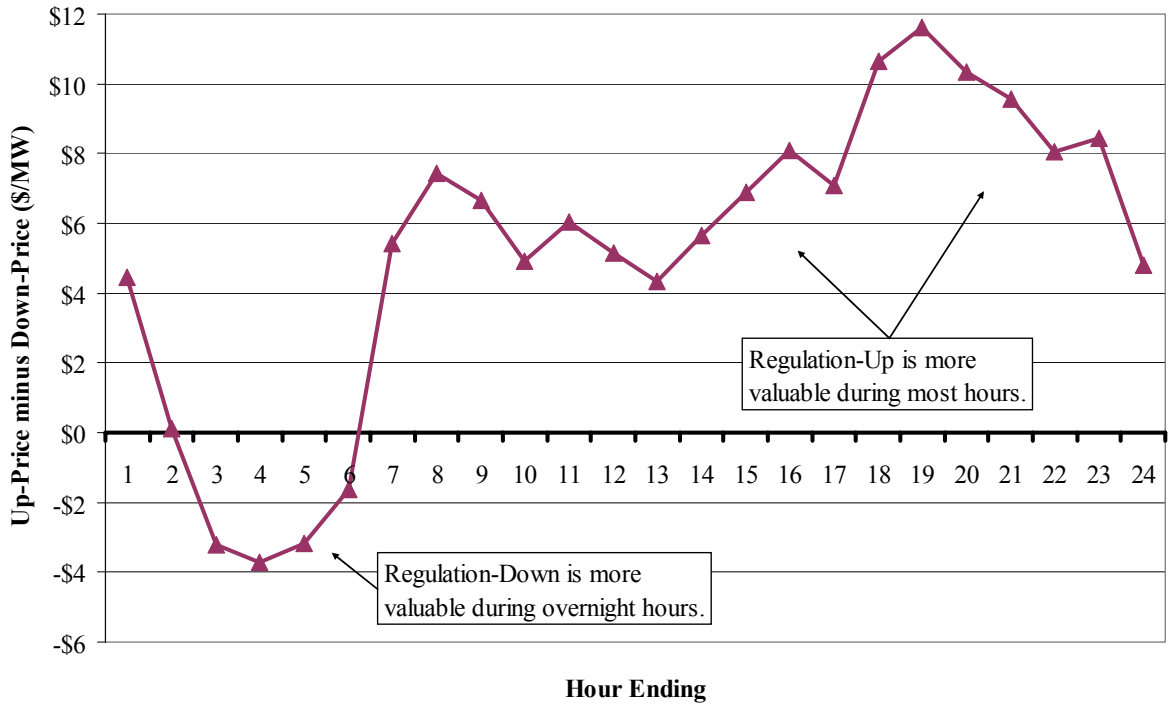


This figure shows that ERCOT requires approximately 1,700 MW of capability prior to the initial ramping period (beginning at 6 AM). Then the requirement jumps up to 2,200 to 2,400 MW during the steepest ramping hours from 6 AM to 9 AM. The requirement declines to 1,800 MW during the late morning and afternoon hours when system load is relatively steady. From 6 PM until midnight, the system is ramping down rapidly and demand for regulation rises to approximately 2,200 MW.

Figure 15 indicates that average regulation prices are closely correlated with the regulation quantities purchased. During non-ramping hours, such as overnight, late morning, and the afternoon, regulation prices average from $8 to $9 per MW. During the ramping hours in early morning and evening, regulation prices average from $12 to $15 per MW. Regulation prices are particularly high during hours ending 6, 7, 23, and 24. This is likely due to the fact that the probability of regulation deployments is much higher in these hours at the beginning and end of the 16 hour bilateral contract period. While up and down regulation are relatively close substitutes and are generally supplied from the same resources, ERCOT runs separate regulation markets reflecting the fact that the marginal costs of providing up and down regulation can differ substantially.

Our next analysis examines the differences between up and down regulation prices. Figure 16 shows the average up regulation price minus the average down regulation price in each hour of the day.

Figure 16: Comparison of Up Regulation and Down Regulation Prices
Up-Price minus Down-Price -- 2003



The figure reveals a distinct intertemporal variation in the price differences. The opportunity costs associated with providing regulation helps explain the inter-temporal pattern of regulation prices. Down regulation prices tend to rise during the off-peak hours—when energy prices are low and there is greater risk that cost will exceed price when a generator is operating above its minimum output level. This is because suppliers of down regulation must operate sufficiently above minimum output levels so they have the ability to reduce output when called on to regulate down in real time. In addition, the overall supply of down regulation is lowest in the early morning hours because fewer units are online and they are operating at relatively low operating levels. Alternatively, up regulation is most expensive during the peak hours when the potential opportunity costs of not producing energy are the highest.

2. Provision of Ancillary Services

To better understand the reserves prices and evaluate the performance of the ancillary services markets, we analyze the capability and offers of ancillary services in this section. The analysis is shown in Figure 17.

Figure 17: Reserves and Regulation Capacity, Offers, and Schedules 2002 & 2003

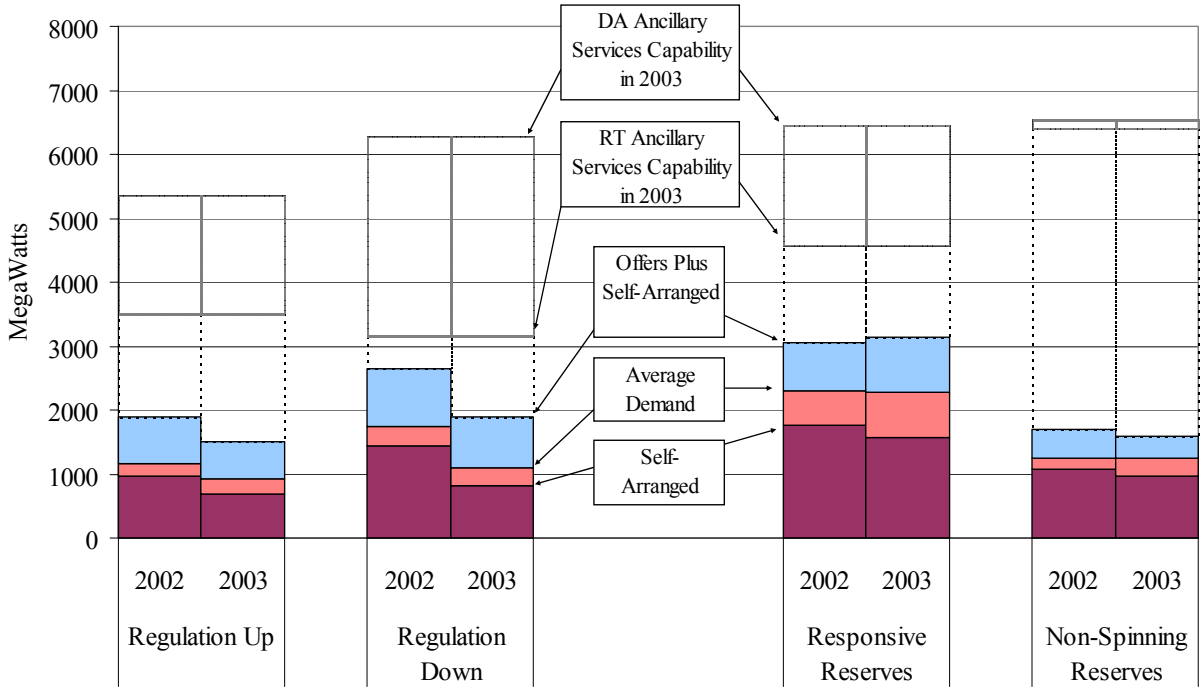


Figure 17 summarizes the quantities of ancillary services offered and self-arranged relative to the total capability and the typical demand for each service. The bottom segment of each bar in the figure is the average quantity of ancillary services self-arranged by owners of resources or through bilateral contracts. The second segment of each bar is the average amount offered and cleared in the ancillary services market. Hence, the sum of the first two segments is the average demand for the service.

The third segment of each bar is the quantity offered into the auction market that is not cleared. Therefore, the sum of the second and third segments is the total quantities offered in each ancillary services auction on average, including the quantities cleared and not-cleared. The empty segments correspond to the ancillary services capability that is not scheduled or offered in

the ERCOT markets. The lower part of the empty segments correspond to the amount of real-time capability that is not offered while the top part of the empty segments correspond to the additional quantity available in the day-ahead that was not offered. Capabilities are generally lower in the real-time because offline units that require significant advance notice to start-up will not be capable of providing responsive reserves or regulation in real time (only capability held on online resources is counted).

The capability shown in Figure 17 incorporates ERCOT's requirements and restrictions for each service type. For regulation, the capability is calculated based on the amount a unit can ramp in five minutes for those units that have the necessary equipment to receive automatic generation control signals on a continuous basis. For responsive reserves, the capability is calculated based on the amount a unit can ramp in ten minutes. This is limited by an ERCOT requirement that no more than 20 percent of the capacity of a particular resource is allowed to provide responsive reserves. However, the responsive reserve capability shown in Figure 17 is not reduced to account for energy produced from each unit, which causes the capability on some resources to be overstated in some hours. Responsive reserves also include approximately 450 MW of Loads Acting as Resources, which are typically offered and procured in the responsive reserves market.

For non-spinning reserves, the figure includes the capability of units able to ramp-up in thirty minutes and able to start-up on short notice. However, it should be noted that any on-line resource with available capacity can provide non-spinning reserves, so the actual capability is larger than shown in the figure. The total capability shown in this figure does not account for capacity of online resources. Hence, the capability that is actually available from a unit in a given hour will generally be less than the amounts shown in this figure because a portion will be used to produce energy.

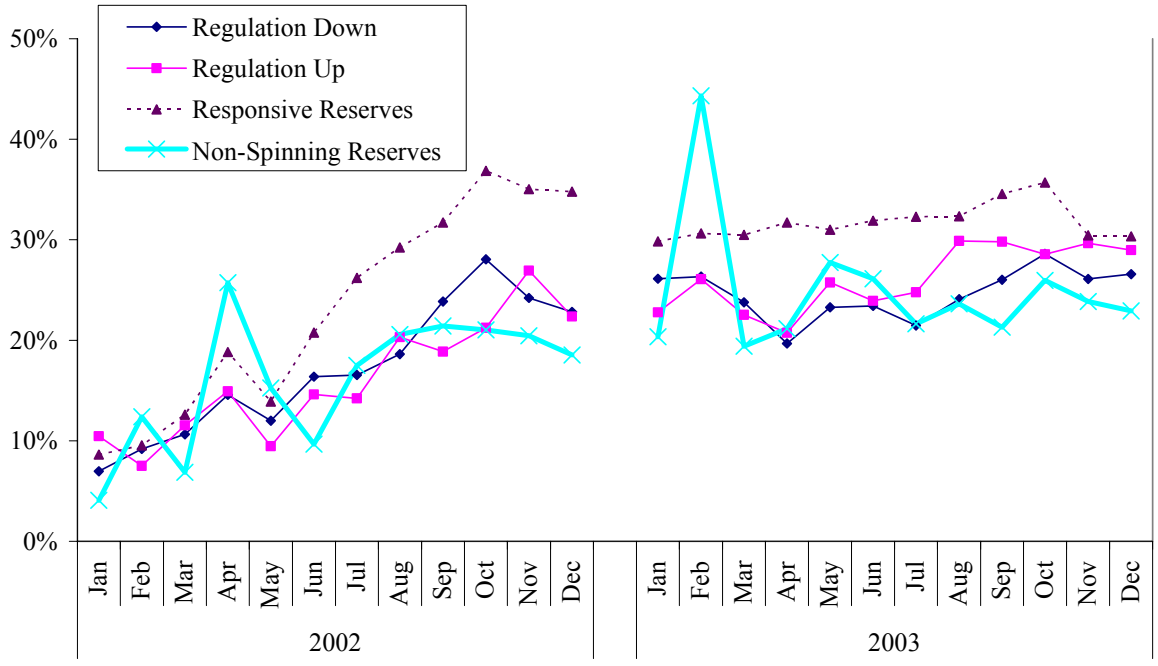
Figure 17 shows that less than one-half of the regulation capability was offered in the regulation market in 2003, while the reserves capability was offered in ranges from 15 percent to 40 percent of the total capability for these services. One explanation for these levels of offers is that the ancillary services markets are conducted ahead of real time so participants may not offer resources that they expect to dispatch to serve their load or to support sales in the balancing energy market. In other words, some of the available reserves and regulation capability becomes

unavailable in real time because the resources are dispatched to provide energy. The current market design creates risk and uncertainty for suppliers that must predict one day in advance whether their resources will be more valuable as energy or as ancillary services.

In addition, participants may not offer the capability of resources they do not expect to commit for the following day. This explanation is less likely because suppliers could submit offer prices high enough to ensure that their costs of committing additional resources to support the ancillary services offers are covered. Nonetheless, there is a substantial quantity of reserves that remain available in real time, but are not offered by the suppliers. This is surprising given the relatively high prices for operating reserves in ERCOT. It is possible that some of the ancillary services capability is withheld in an attempt to increase the ancillary services clearing prices. The analysis in this section is not sufficient to make that determination given that there are multiple factors that may be contributing to these offer patterns.

Finally, Figure 17 shows that a relatively high share of these services is self-supplied. These services can be self-supplied from owned resources or from resources purchased bilaterally. To evaluate the quantities of ancillary services that are not self-supplied more closely, Figure 18 shows the share of each type of ancillary service that is purchased through the ERCOT market.

Figure 18: Portion of Reserves and Regulation Procured Through ERCOT 2002 & 2003

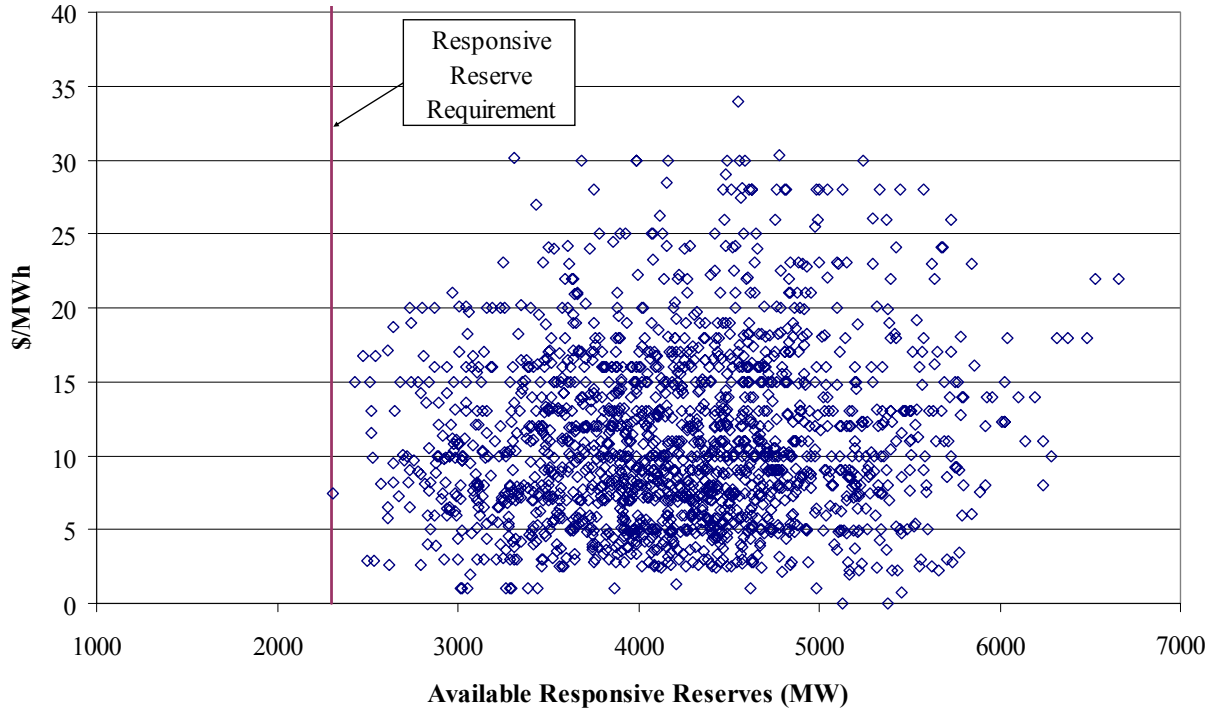


This figure shows that purchases of all ancillary services from the ERCOT markets increased substantially during 2002 and remained at the higher levels in 2003. Hence the portion of the ancillary services requirements procured through the ERCOT-administered market has not increased substantially compared to the latter half of 2002. However, we would expect that if higher portions of the available responsive reserves and regulation capability are offered into the market in the future, thereby increasing the market's liquidity and reducing ancillary services prices, then a higher share of the requirements would be met through the ERCOT market. We do not anticipate a significant change in the offer patterns under current markets. Jointly optimizing the reserves and energy markets in the context of the Texas Nodal markets could serve to increase the liquidity of these markets by reducing the economic costs of selling ancillary services under the current sequential market design.

The final analysis in this section evaluates the prices prevailing in the responsive reserves market during 2003. Prices in this market are significantly higher than in other markets that co-optimize the dispatch of energy and responsive reserves. Lower prices occur in co-optimized markets because in most hours there is substantial excess online capacity that can provide responsive reserves at very low incremental costs. For example, a steam unit that is not economic to operate at its full output in all hours will have output segments that can provide responsive reserves at very low incremental costs. If the surplus responsive reserves capability from online resources is relatively large in some hours, one can gauge the efficiency of the ERCOT reserves market by evaluating the prices in these hours.

Hence, Figure 19 plots the hourly real-time responsive reserves capability against the responsive reserves prices in the peak afternoon hours (2 PM to 6 PM). The capability calculated for this analysis reflects the actual energy output of each generating unit. Hence, units producing energy at their maximum capability will have no available responsive reserves capability. The figure also shows the responsive reserves requirement of 2,300 MW to show the amount of the surplus in each hour.

Figure 19: Hourly Responsive Reserves Capability vs. Market Clearing Price Afternoon Peak Hours -- 2003



This figure indicates a somewhat random pattern of responsive reserves prices in relation to the hourly available responsive reserves capability in real time. Although not obvious from the scatter plot, prices are negatively correlated to the responsive reserves capability, as expected. However, the dispersion of prices is wide. Particularly surprising is the frequency with which the price exceeds $10 per MWh when the available responsive reserves capability is more than 2,000 MW higher than the requirement. In these hours, the marginal costs of supplying responsive reserves should be zero. These results reinforce the potential benefits promised by jointly optimizing the operating reserves and energy markets, which we would recommend in the context of the nodal markets currently under consideration.

Non-spinning reserves are purchased on an as-needed basis whenever ERCOT predicts a balancing energy shortage at least one hour in advance. Non-spinning reserves are resources that can be brought on-line within 30 minutes. Thus, off-line quick-start units can provide non-spinning reserves. In addition, any resource that plans to be on-line with capacity not already scheduled for energy, regulation, or responsive reserves can also provide non-spinning reserves.

Figure 20 shows the relationship between excess available non-spinning reserves capability and the market clearing price in the non-spinning reserves auction for the afternoon hours in 2003.

Figure 20: Hourly Non-Spinning Reserves Capability vs. Market Clearing Price
Afternoon Peak Hours – 2003

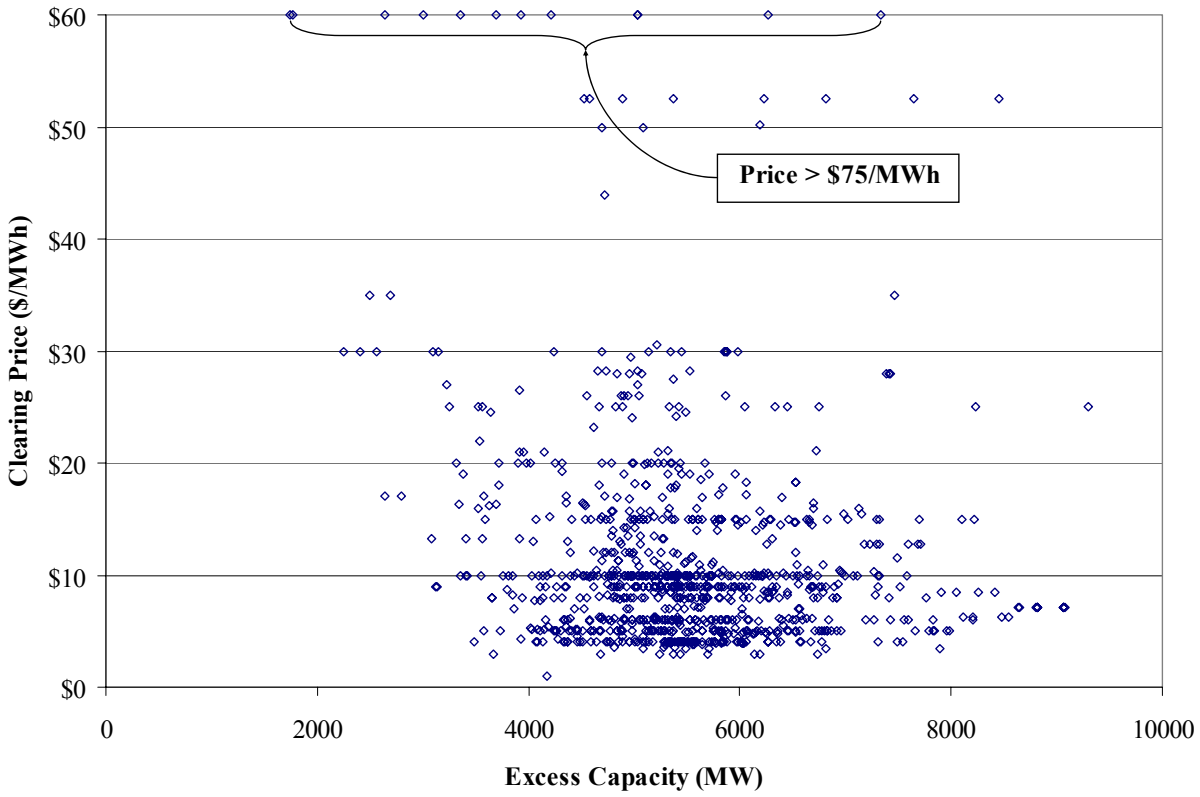


Figure 20 shows that there were at least 2,000 MW of excess capacity capable of providing non-spinning reserves in every hour when it was purchased. Like the results for responsive reserves, these prices have only a slightly negative correlation to the non-spinning reserves capability and there is a wide dispersion in prices even in hours with similar excess capability. Again, the lack of co-optimized markets for energy, regulation, and reserves may be a primary contributing factor to the high prices for non-spinning reserves when there are large quantities of excess capacity available.

C. Net Revenue Analysis

Net revenues from the energy, operating reserves, and regulation markets together provide the economic signals that govern suppliers' decisions to invest in new generation or retire existing generation. Net revenue is defined as the total revenue that can be earned by a generating unit

less its variable production costs. Hence, it is the revenue in excess of short-run operating costs and is available to recover a unit's fixed and capital costs.

In a sustainable long-run equilibrium, the markets should provide sufficient net revenue to allow an investor to break-even on an investment in a new generating unit. In the short-run, if the net short-run revenues produced by the market are not sufficient to justify entry, then one or more of three conditions exist:

a. New capacity is not needed because there is sufficient generation already available;

b. Load levels, and thus energy prices, are temporarily below long-run expected levels (this could be due to mild weather or other factors); or

c. Market rules are causing revenues to be reduced inefficiently.

Likewise, the opposite would be true if prices provide excessive net revenues in the short-run. The persistence of excessive net revenues in the presence of a capacity surplus is an indication of competitive issues or market design flaws. In this section, we analyze the net revenues that would have been received in 2002 and 2003 by various types of generators in each of the ERCOT zones.

Figure 21 shows the results of the net revenue analysis for two types of units. The first type is a gas combined-cycle (with an assumed heat rate of 7,000 BTU/kWh). The second type is a gas turbine (with an assumed heat rate of 10,500 BTU/kWh). Net revenue is calculated by assuming the unit will produce energy in any hour for which it is profitable and by assuming it will be available to sell reserves and regulation in other hours. These revenues are reduced based on the assumed outage rate for each unit. For purposes of this analysis, we assume the heat rates cited above for each unit, $4 per MWh variable operating and maintenance costs, and a total outage rate (planned and forced) of 10 percent. Some units, generally those located in unique locations that are used to resolve local transmission constraints, also receive a substantial amount of revenue through uplift payments (i.e., Out-of-Merit Energy, Out-of-Merit Commitment, and Reliability Must Run payments). This source of revenues is assumed to be zero for purposes of our analysis.

Figure 21: Estimated Net Revenue 2002-2003

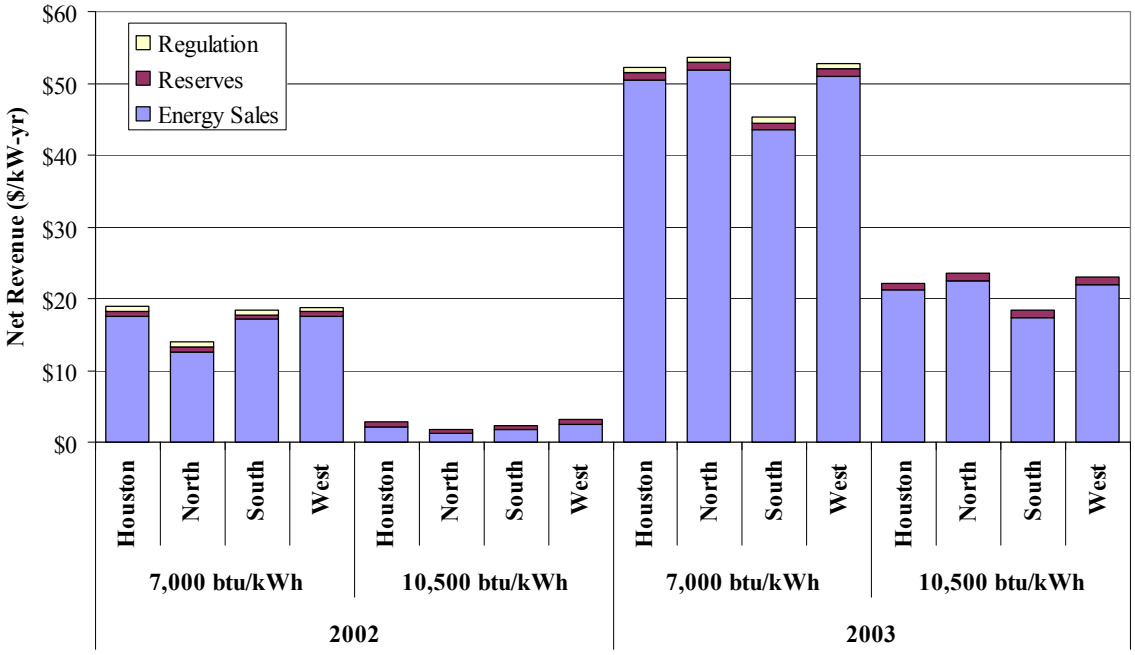


Figure 21 shows that the estimated net revenue was significantly higher in 2003 than in 2002. This is largely because higher natural gas prices in 2003 led to substantially higher energy prices. While the higher natural gas prices also lead to higher costs for these new units, these units are more efficient than the resources that set prices in some hours. Therefore, an increase in natural gas prices can increase the margin for the new units and result in higher net revenue.

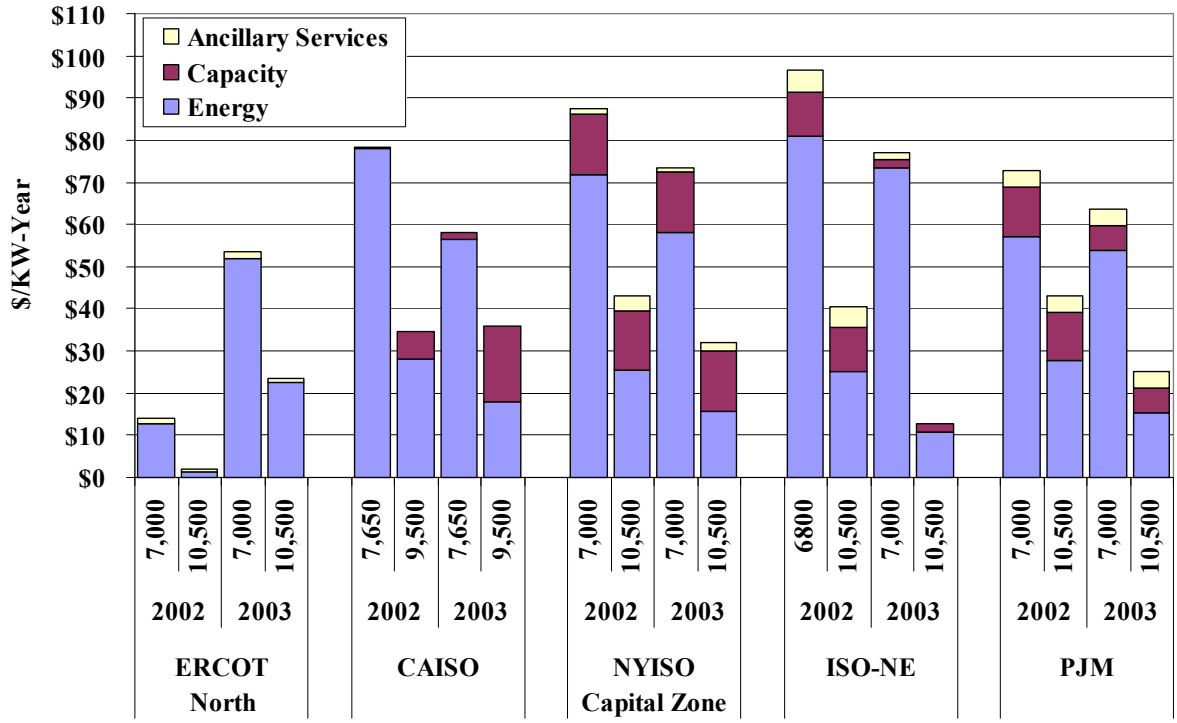
Importantly, there were also a much larger number of relatively high-priced hours in 2003, which contributed to the higher net revenues. These higher prices frequently occurred under moderate load conditions when one would not expect high prices. We have determined that these prices are due to scheduling and ramping issues under the current market design and to the fact that a large share of the available balancing energy capability is not offered in the balancing energy market. These issues are analyzed in Section II below.

Although net revenues were substantially higher in 2003, neither type of new generating unit would have earned sufficient net revenue to make the investment profitable. Based on our estimates of investment data for new units, the net revenue required to satisfy the annual fixed costs (including capital carrying costs) of a new gas turbine unit is $70 to $80 per kW-year. For a new combined cycle unit, net revenue requirements are more that $100 per kW-year. Although

the net revenue increased considerably from 2002 to 2003, it remained less than half of the amount necessary to support new investment. Hence, these hypothetical new units would not be profitable. This is not surprising given the surplus of capacity that currently exists in ERCOT.

We also compared the net revenue in the ERCOT market with net revenue in other centralized wholesale markets. Figure 22 compares estimates of net revenue for each of the auction-based wholesale electricity markets in the U.S.: (a) the ERCOT North Zone, (b) the California ISO, (c) the New York ISO, (d) ISO New England, and (e) the PJM ISO. The figure includes estimates of net revenue from (a) energy, (b) reserves and regulation, and (c) capacity. ERCOT does not have a capacity market, and thus, does not have an estimate for net revenue from capacity sales.[12]

Figure 22: Comparison of Net Revenue between Markets 2002 – 2003



12 The California ISO does not report capacity and ancillary services net revenue separately, so it is shown as a combined block in Figure 22. Generally, estimates were performed for a theoretical new combined-cycle unit with a 7,000 BTU/kWh heat rate and a theoretical new gas turbine with a 10,500 BTU/kWh heat rate. However, the California ISO reports net revenues for 7,650 and 9,500 BTU/kWh units, and, in 2002, the ISO–New England reported net revenues for a 6,800 BTU/kWh combined-cycle unit.

Based on Figure 22, net revenues decreased slightly or remained flat from 2002 to 2003 for every market except ERCOT, where estimated net revenue increased by a factor of three for a theoretical combined-cycle unit and by more than a factor of ten for a theoretical gas turbine. This difference can be explained by a number of factors. First, ERCOT is much more dependent on natural gas than the other markets. The sharp increase in natural gas prices in the other regions does not translate as directly into higher electricity prices because natural gas units are displaced in many hours by other types of units. Second, many of the natural gas units in the Northeast are dual-fueled, allowing them to switch to oil when natural gas becomes relatively expensive. This causes the net revenue for the hypothetical new units that can only burn natural gas to fall.

Third, a substantial amount of new capacity has been installed over the last two years in the Northeast and load levels have been relatively low during 2003 due to extremely mild weather. These factors also contribute to lower net revenue in the Northeast. Finally, the increased frequency of relatively high electricity prices in ERCOT, as discussed above, also contributed to the increase in net revenue. The sources of these increases are evaluated in Section II.

Despite increases in the estimates of net revenue for ERCOT from 2002 to 2003, they are still lower than in the other markets, with the exception of a new gas turbine in New England. However, if net revenue from capacity markets is excluded, ERCOT would actually show higher net revenues for a new gas turbine than New York or PJM. None of these markets produces net revenue close to the amounts needed to support investment in such resources.

II. FORWARD SCHEDULING AND RESOURCE PLANS

The ERCOT market protocols require QSEs to submit balanced load and energy schedules prior to the operating hour. These forward schedules are initially submitted in the day ahead and can be subsequently updated during the adjustment period up until one hour before the operating hour. QSEs are also required to submit a resource plan that indicates, among other things, units that are expected to be online and producing energy. Under ERCOT's relaxed balanced schedules policy, the load schedule is not required to approximate the QSE's projected load. When a QSE's forward schedule is less than its actual real-time load, its generation is under-scheduled and it will purchase its remaining energy requirements in the balancing energy market at the balancing energy price. Likewise, when a QSE's forward schedule is greater than actual load, its generation is over-scheduled and it will sell the residual in the balancing market at the balancing energy price.

The QSE schedules and resource plans are the fundamental sources of information to ERCOT regarding the supply and demand components of the ERCOT market. In this section, we evaluate certain aspects of the QSE schedules and resource plans and we draw conclusions about balancing energy prices, market participants' behavior, and the efficiency of the market design. The results of this analysis lead us to make several recommendations to improve the operation of the current markets.

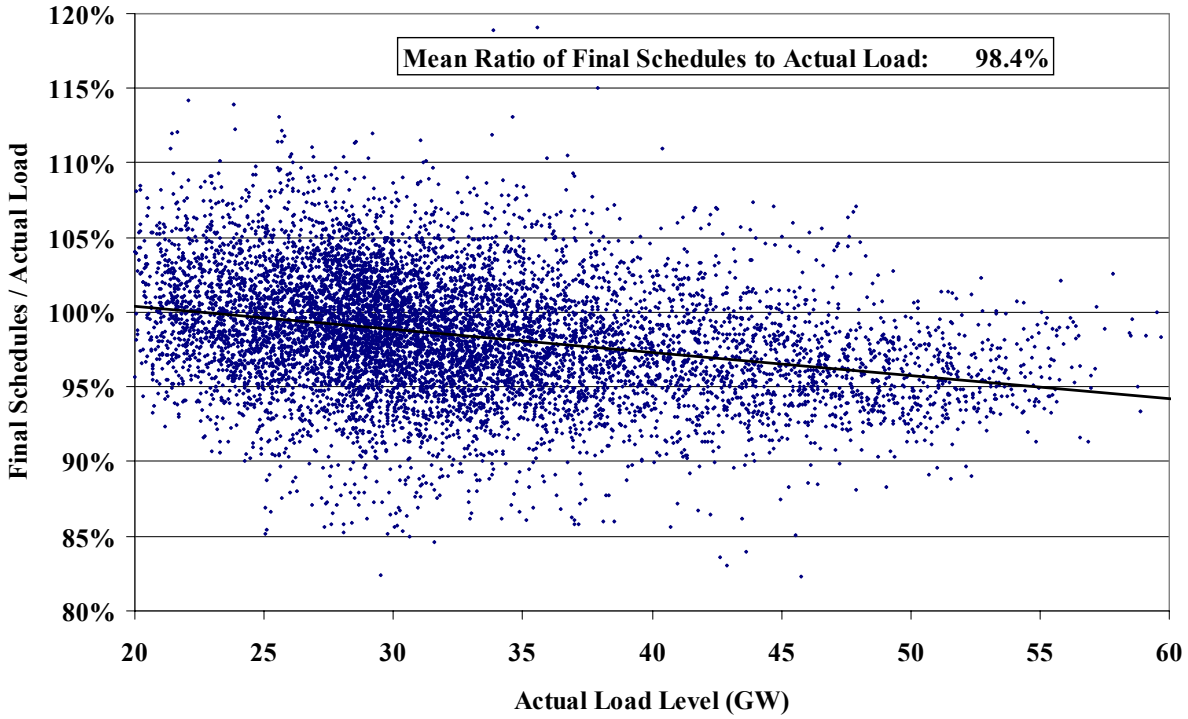
This section analyzes a number of issues, beginning with forward scheduling by QSEs. The analysis focuses on the degree to which forward schedules depart from actual load levels. Our second analysis focuses on the balancing energy market and, in particular, how scheduling patterns affect balancing energy deployments and prices. The third analysis evaluates the rate of participation in the balancing energy market. Finally, we analyze market participant resource plans to determine whether the information provided to ERCOT regarding generating units' projected commitment and output levels is affected by certain adverse incentives embodied in the ERCOT protocols.

A. Forward Scheduling

In this subsection, we evaluate forward scheduling patterns by comparing forward schedules to actual real-time load. We focus on the forward schedules at two points in time. First, we will refer to the final schedule, which is the last schedule submitted by the QSE prior to the operating hour. Second, we will refer to the day-ahead schedule, submitted by the QSE in the day ahead.

To provide an overview of the scheduling patterns, Figure 23 shows a scatter diagram that plots the ratio of the final schedules to the actual load level during 2003.

Figure 23: Ratio of Final Schedules to Actual Load
All ERCOT – 2003



The ratio shown in Figure 23 will be greater than 1 when the final schedule is greater than the actual load. Therefore, in general, the figure shows that load is slightly under-scheduled in the aggregate, as indicated by an average ratio of the final schedules to actual load of 98.4 percent. The figure also includes a trend line indicating that the ratio of final schedules to actual load tends to decrease as load rises. In particular, the ratio is 100 percent at 23 GW and declines to 95 percent at just below 55 GW.

This result is counter-intuitive. Normally, one would expect balancing energy prices to be more volatile at high load levels. Therefore, if market participants were generally risk averse, they would be expected to schedule forward to cover their load during high load periods rather than reducing their forward scheduling levels during those periods. There may be several explanations for this scheduling pattern. First, while the data suggests that QSEs rely more on the balancing energy market at higher load levels, doing so does not necessarily subject them to the attendant price risk. Financial contracts or derivatives may be in place that protect market participants from the price risk in the balancing energy market, such as a contract for differences. Second, they can cover themselves by bidding enough generation to cover their load needs in the balancing market. Last, the fact that balancing energy prices have not risen predictably with actual load levels (as shown above) may provide an incentive for some market participants to purchase peak energy from the balancing market that they need to satisfy their load.

Figure 24 is a further analysis of final schedules that shows the ratio of final schedules to actual load evaluated at five different load levels for each of the ERCOT zones.

Figure 24: Average Ratio of Final Schedules to Actual Load by Load Level All Zones – 2003

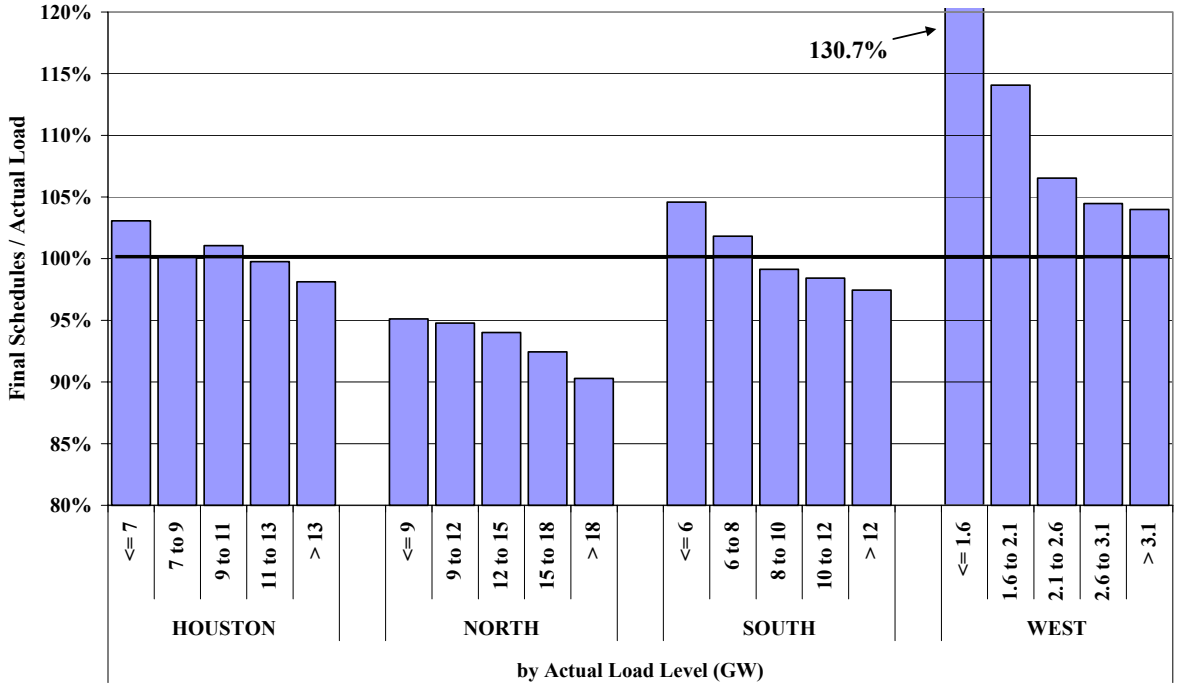


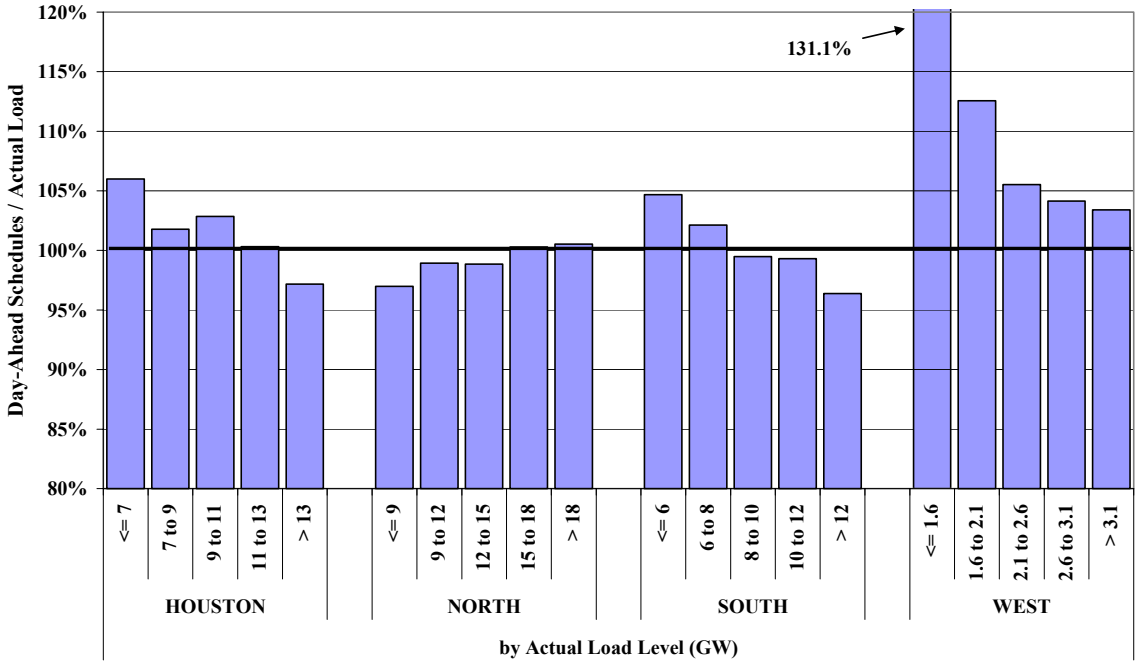
Figure 24 shows that relative to actual load, the final schedule quantity decreases in each of the four zones as actual load increases, which is consistent with Figure 23. The Houston Zone and

South Zone are generally scheduled at nearly 100 percent of the actual load over the various load levels, although the ratios decline as load increases. The North Zone is substantially under-scheduled on average at each load level, ranging from 5 percent under-scheduled when North Zone load is below 9 GW to 10 percent under-scheduled when load is greater than 18 GW. The West Zone is consistently over-scheduled, although the amount of over-scheduling decreases as load increases -- from 31 percent over-scheduled at loads below 1.6 GW to 4 percent over-scheduled at load levels above 3.1 GW.

The result of these scheduling patterns is that the QSEs in the North Zone are net buyers of balancing energy to the extent that they do not offer generation in the balancing energy market to cover their deficits. In contrast, QSEs in the West Zone are net sellers of balancing energy since they must sell a surplus of 4 to 31 percent of their load on average. Persistent load imbalances are not necessarily a problem. Later in this section, we perform a resource-level analysis to identify at least one factor that may contribute to this scheduling pattern.

We next evaluate the day-ahead schedules relative to actual load in Figure 25. The figure is analogous to Figure 24. It shows the ratio of day-ahead schedules to actual load by load level for each of the four zones in ERCOT. The load levels are divided into five roughly equal groups.

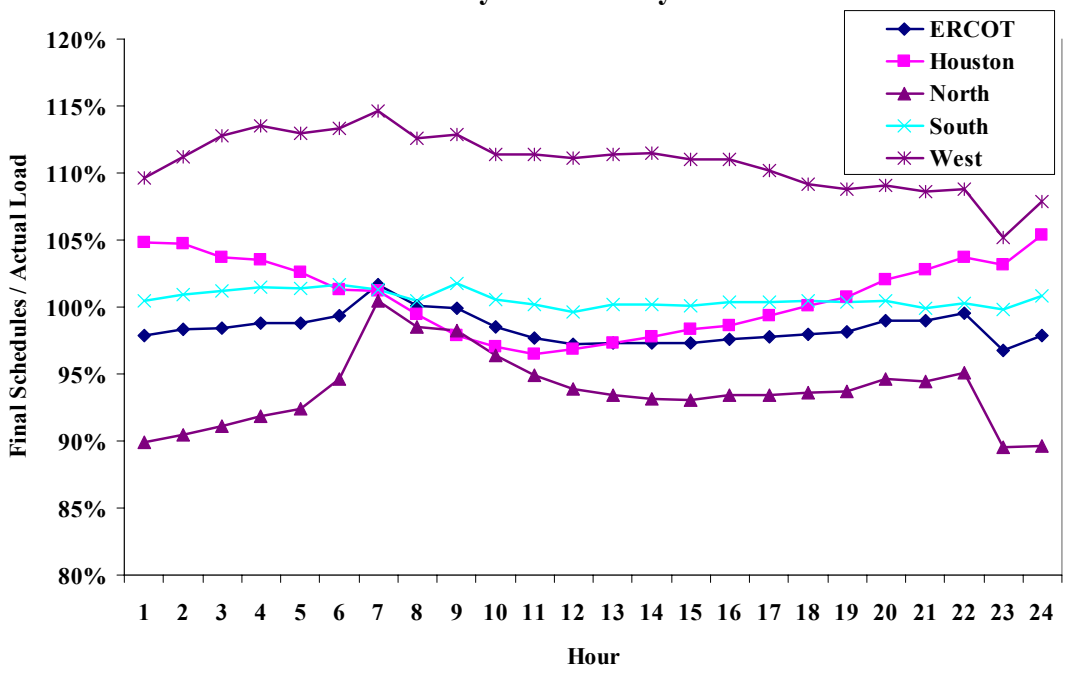
Figure 25: Average Ratio of Day-Ahead Schedules to Actual Load by Load Level
All Zones – 2003



For three of the four zones, Figure 25 shows day-ahead scheduling results that are comparable, both in magnitude and pattern, to the scheduling levels shown in Figure 24 for final load schedules. Day-ahead load schedules in the Houston, South, and West Zones are all negatively correlated with actual load levels, with the day-ahead schedules in the Houston and South Zones averaging close to 100 percent of actual load. However, scheduling in the North Zone is very different in the day-ahead than in real time. The ratio of day-ahead schedules to actual load increases as actual load increases while the ratio of final schedules to actual load decreases as load increases.

Additionally, the ratio of day-ahead scheduled load to actual load in the North Zone is much closer to 100 percent than the ratio of the final schedules to actual load. At the highest load levels, the QSEs in the North Zone submit day-ahead load schedules averaging 101 percent of actual load, while final load schedules average only 90 percent of actual real-time load. Although there is no obvious explanation for this scheduling pattern in the North Zone, the more detailed analysis of out-of-merit commitment and dispatch described below provides some insight. To further analyze forward scheduling, Figure 26 shows the ratio of final schedules to actual load by hour-of-day for each of the four zones in ERCOT as well as for ERCOT as a whole.

Figure 26: Average Ratio of Final Schedules to Actual Load All Zones by Hour of Day - 2003



This figure shows that on an ERCOT-wide basis, final schedules are close to actual load (between 98 percent and 100 percent) during hours ending 1 to 6. At hour ending 7, the ratio rises to 102 percent, the highest of all hours. In the remaining hours, the ratio declines to between 96 percent and 100 percent.

The scheduling in the North Zone heavily influences ERCOT's overall scheduling pattern. The average fraction of load scheduled before real time for the North Zone is shown between 90 percent and 95 percent during hours ending 1 to 6. The highest schedule-to-actual-load ratio is reached in hour ending 7, when approximately 100 percent of actual load is scheduled before real-time. The ratio decreases to between 93 percent and 94 percent in the afternoon hours. In hours ending 23 and 24, the scheduled quantities drop to just below 90 percent of actual load.

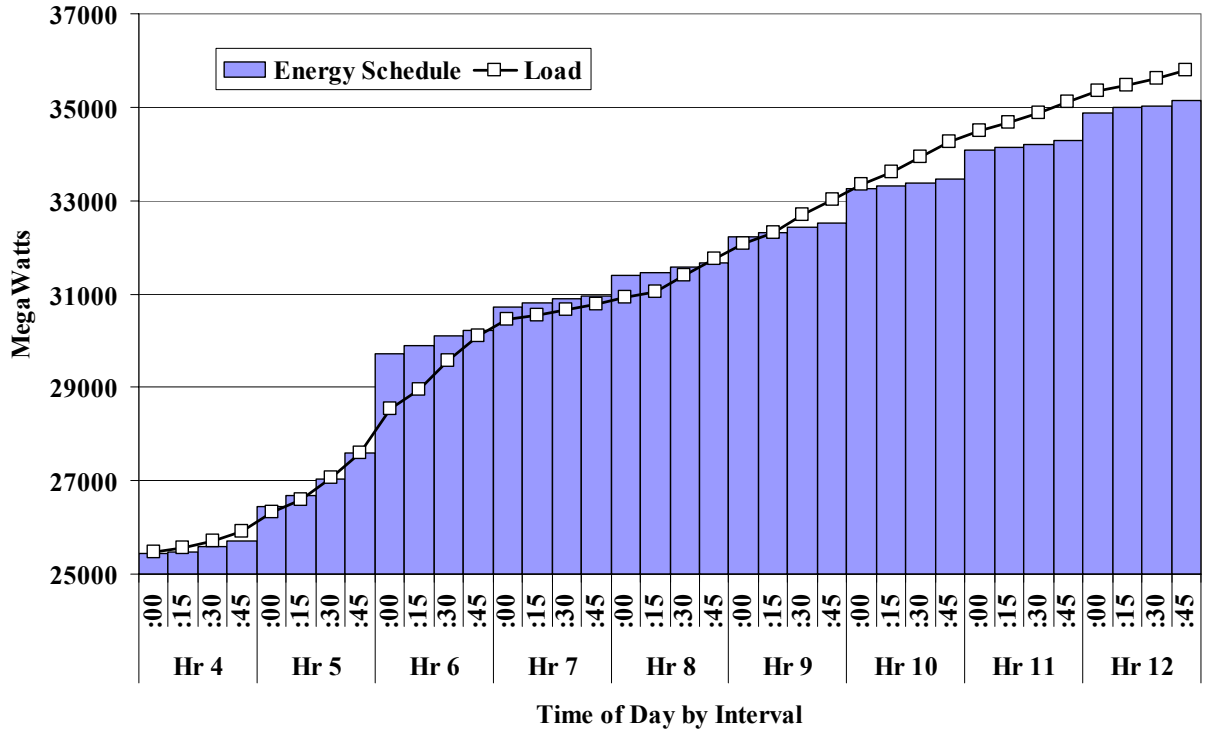
Hour ending 7 and hour ending 22 represent start and end points of the 16 hour block of peak hours commonly used in bilateral contracts. Hence, a logical explanation for the patterns shown in Figure 26 is that participants tend to submit schedules consistent with their bilateral transaction positions. This is not irrational if the market participants also submit balancing energy offers to optimize the energy that is actually deployed. In addition, this pattern of forward scheduling is consistent with the notion that market participants bear additional price risk in ramping hours (as shown in the prior section), accounting for their propensity to schedule a larger portion of their needs during these periods. However, the latter explanation does not explain the sharp change in scheduling in hours ending 7 and 22.

B. Balancing Energy Market Scheduling

In the previous section, we analyzed balancing energy prices and load and found that actual load is not the primary determinant of balancing energy prices. In this section, we investigate whether balancing energy prices are influenced by market participants' scheduling practices that tend to intensify the demand for balancing energy during hours when load is ramping.

We begin our analysis by examining factors that determine the demand for balancing energy during periods when load is ramping up and periods when it is ramping down. Figure 27 shows average energy schedules and actual load for each interval from 4 AM to 1 PM during 2003. In general, energy schedules that are less than the actual load result in balancing energy purchases while energy schedules higher than actual load result in balancing energy sales.

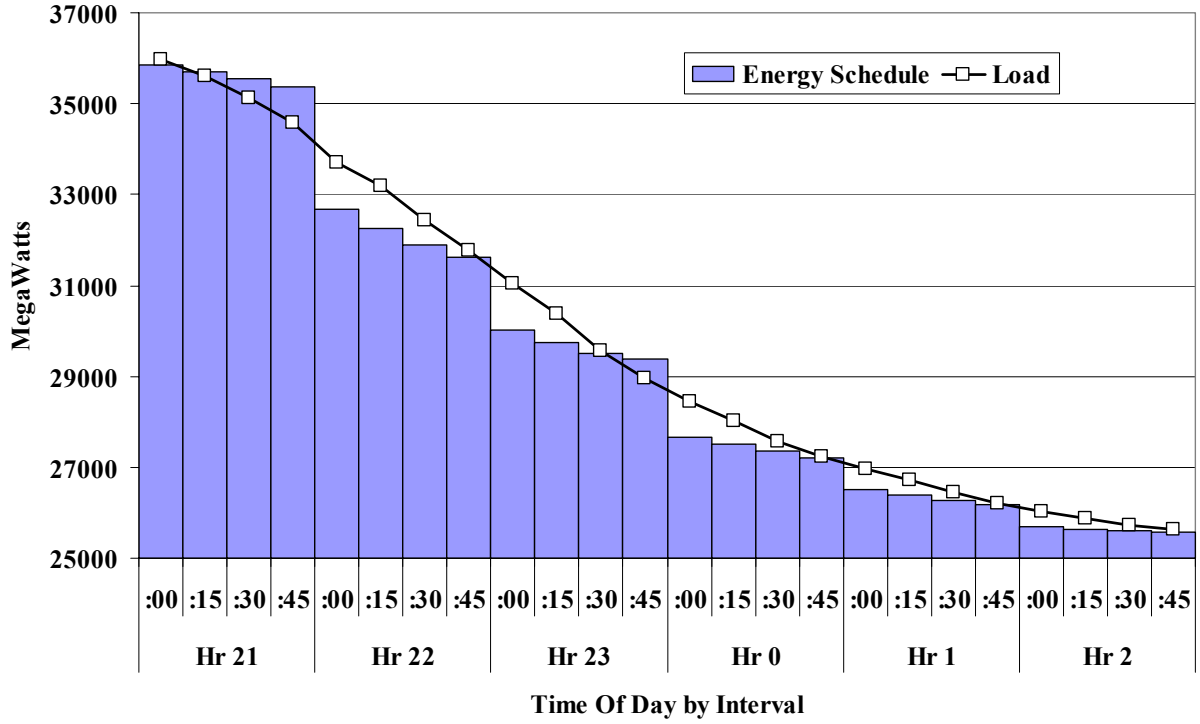
Figure 27: Final Schedules during Ramping-Up Hours
2003



On average, load increases from less than 26 GW to nearly 36 GW in the nine hours shown in Figure 27. The average increase per 15-minute interval is approximately 290 MW, although the rate of increase is greatest from 5:45 AM to 7:00 AM and relatively flat from 7:00 AM to 8:30 AM. The progression of load during ramping-up hours is steady relative to the progression of energy schedules. Energy schedules rise less smoothly, with small increases from the first to fourth interval in each hour and large increases from the fourth interval to the first interval of the next hour. For instance, the average energy schedule increases approximately 2 GW from the last interval of the hour ending 6 AM to the interval beginning at 6 AM, while the average energy schedule increases by several hundred megawatts in the subsequent three intervals.

The same scheduling patterns exist in the ramping-down hours. Figure 28 shows average energy schedules and load for each interval from 9 PM to 2 AM during 2003.

Figure 28: Final Schedules during Ramping-Down Hours
2003



On average, load drops from approximately 36 GW to less than 26 GW in the six hours shown in Figure 28. The average decrease per 15-minute interval is approximately 420 MW, although the rate of decrease is greatest from 9:45 PM to midnight. The progression of load during ramping-down hours is steady relative to the progression of energy schedules. As during the ramping-up hours, energy schedules decrease in relatively large steps at the top of each hour. For instance, the average energy schedule drops almost 3 GW from the last interval of the hour ending at 9 PM to the interval beginning at 10 PM.

The sudden changes in energy schedules that occur at the beginning of each hour during ramping-up hours and at the end of each hour during ramping-down hours arise from the fact that one-third of the load in ERCOT is scheduled by QSEs that submit energy schedules that change hourly, while the other two-thirds of the load is scheduled by QSEs that submit energy schedules that change every 15 minutes. Deviations between the energy schedules and actual loads will result in purchases or sales in the balancing energy market. Specifically, net balancing up energy equals real-time load minus scheduled energy. Hence, Figure 27 indicates that during ramping-up hours, QSEs tend to purchase balancing energy on net at the end of each hour and sell

balancing energy at the beginning of each hour. On the other hand, Figure 28 indicates that during ramping-down hours, QSEs tend to sell balancing energy on net at the beginning of each hour and purchase balancing energy at the end of each hour.

To evaluate the effects of systematic over- and under-scheduling more closely, we analyzed balancing energy prices and deployments in each interval during the ramping-up period and ramping-down period (consistent with the periods shown in Figure 27 and Figure 28). This analysis is similar to that shown in Figure 11 and Figure 12, except instead of showing balancing energy prices relative to load, we show balancing energy prices relative to balancing energy deployments. Figure 29 shows the analysis for the ramping-up hours.

Figure 29: Balancing Energy Prices and Volumes
Ramping-Up Hours -- 2003

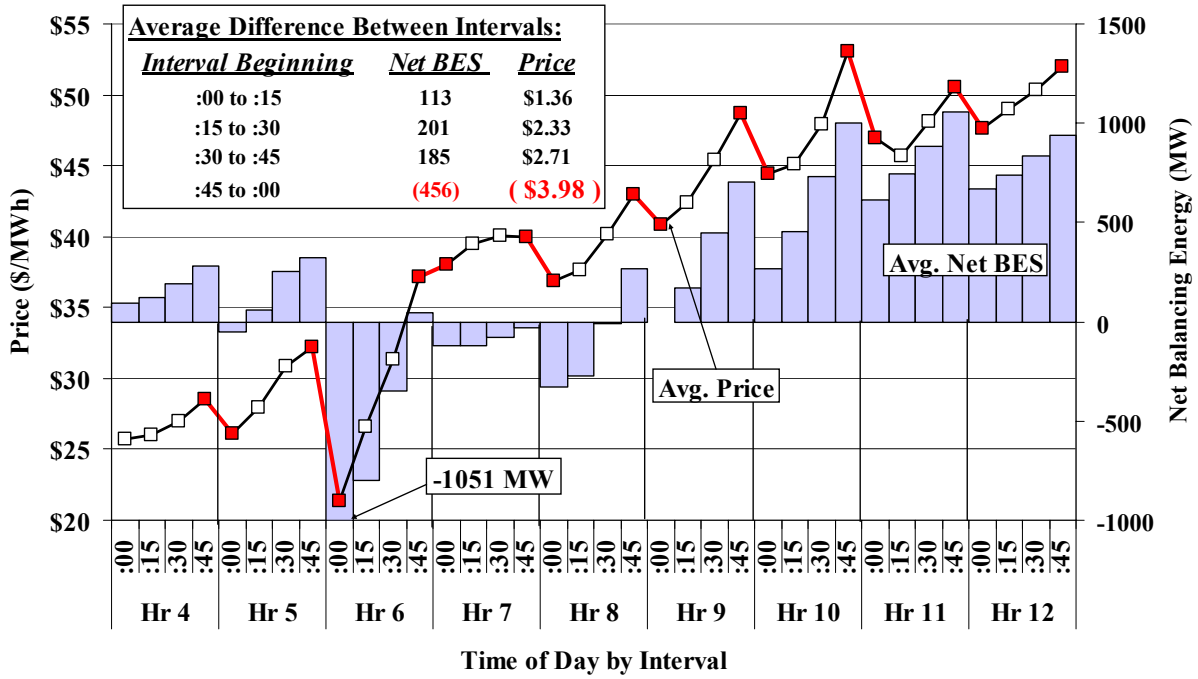
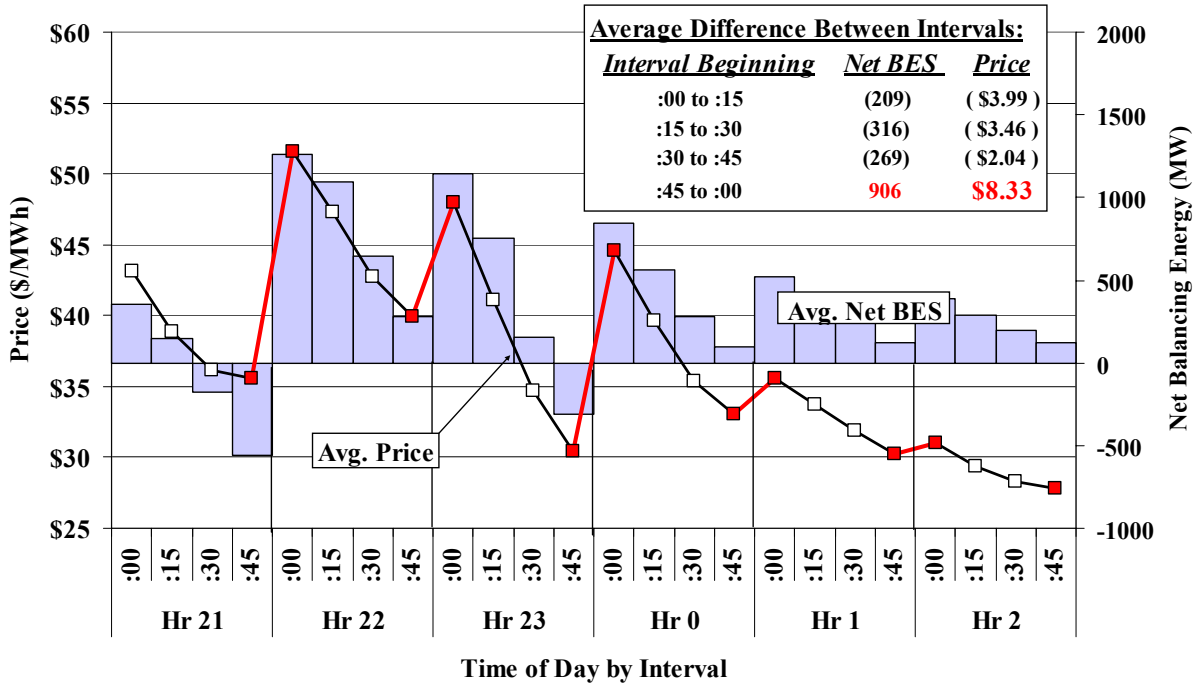


Figure 29 reveals two key aspects of the balancing energy market. First, as discussed above, balancing energy prices are highly correlated with balancing energy deployments. Second, there is a distinct pattern of increasing purchases during the hour. At the beginning of the hour, purchases tend to be smaller than at the end of the hour. This is consistent with the notion that hourly schedules are established at a level that corresponds to an average expected load for the hour. Whatever the reason for the scheduling patterns that create these balancing deployments,

the effect on the ERCOT prices is inefficient. These prices are relatively volatile and could result in erratic dispatch signals to the generators. Figure 30 shows the same analysis for the ramping-down hours.

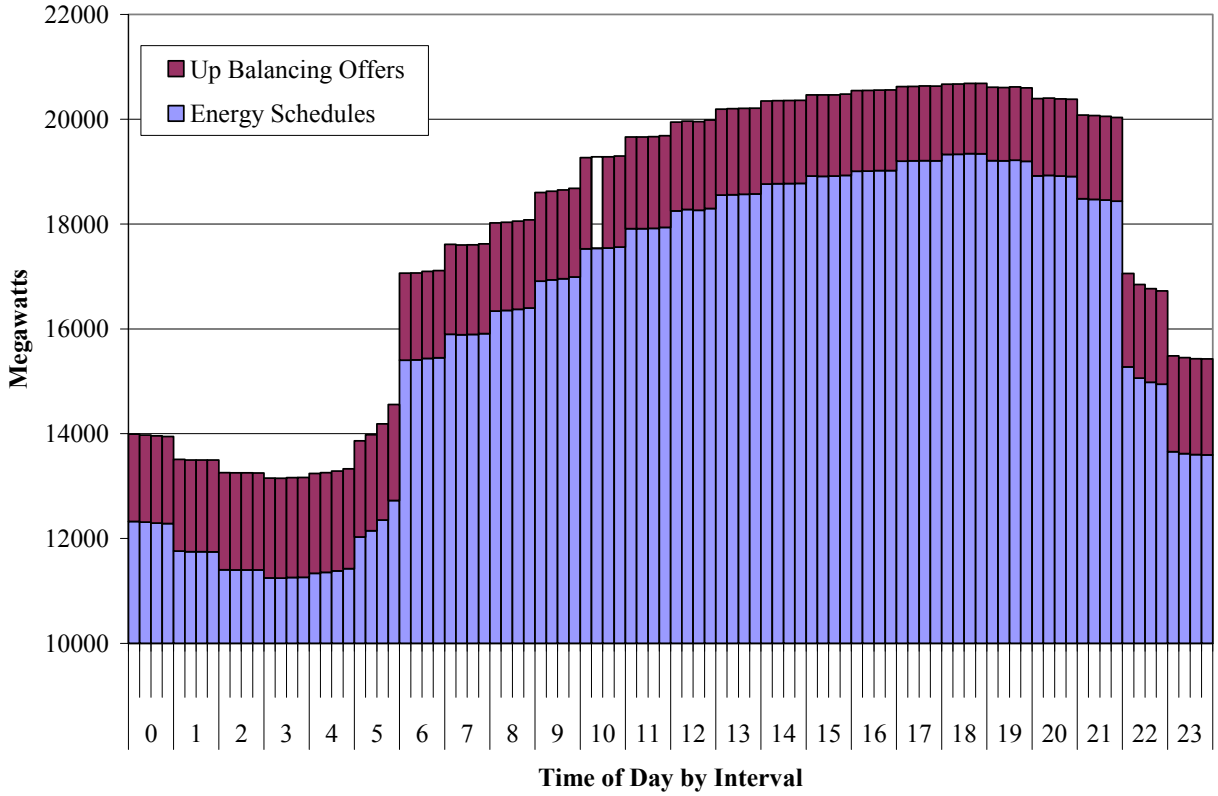
Figure 30: Balancing Energy Prices and Volumes
Ramping-Down Hours -- 2003



During ramping down hours, at the beginning of the hour, actual load tends to be higher than energy schedules, resulting in substantial balancing energy purchases. At the end of the hour actual load tends to be lower relative to the energy schedules, resulting in lower balancing energy demand.

While QSEs have the option to submit flexible schedules (i.e., every 15 minutes), most of the QSEs schedule only on an hourly basis, making little or no changes on a 15-minute basis. However, the two largest suppliers in ERCOT tend to schedule much more flexibly than other QSEs. Our next step is to analyze the scheduling patterns of the two largest QSEs compared to all other QSEs by interval over the entire day. Figure 31 shows the average quantity of energy schedules and balancing-up offers in 2003 for all QSEs in ERCOT except the largest two.

Figure 31: Final Energy Schedules and Balancing-up Offers
All QSEs except the Largest Two



This figure shows that there is almost no change in the energy schedules on a 15-minute basis, but relatively large changes from hour to hour. It is primarily the scheduling patterns by these QSEs that result in the balancing deployments and prices shown in Figure 29 and Figure 30. In addition to the fact that these QSEs generally schedule hourly, this figure shows the sharp schedule changes that occur at the beginning and end of the 16 peak hours commonly used in bilateral contracts (hour ending 7 to hour ending 22). Energy schedules increase by 2675 MW on average from the last interval of hour ending 6 to the first interval of hour ending 7, an increase of more than 20 percent in just one interval.

The scheduled energy decreases even more abruptly at the end of the peak bilateral contract period. Scheduled energy decreases by 3170 MW on average from the last interval of hour ending 22 to the first interval of hour ending 23. The next two hours also show relatively large decreases in scheduled energy of 1290 MW and 1270 MW, respectively. These large hourly changes in energy schedules are a primary determinant of the balancing energy price fluctuations shown in this section.

Figure 32 shows the energy scheduling patterns of the two largest QSEs, which account for approximately one half of the energy schedules.

Figure 32: Final Energy Schedules and Balancing up Offers
Largest Two QSEs

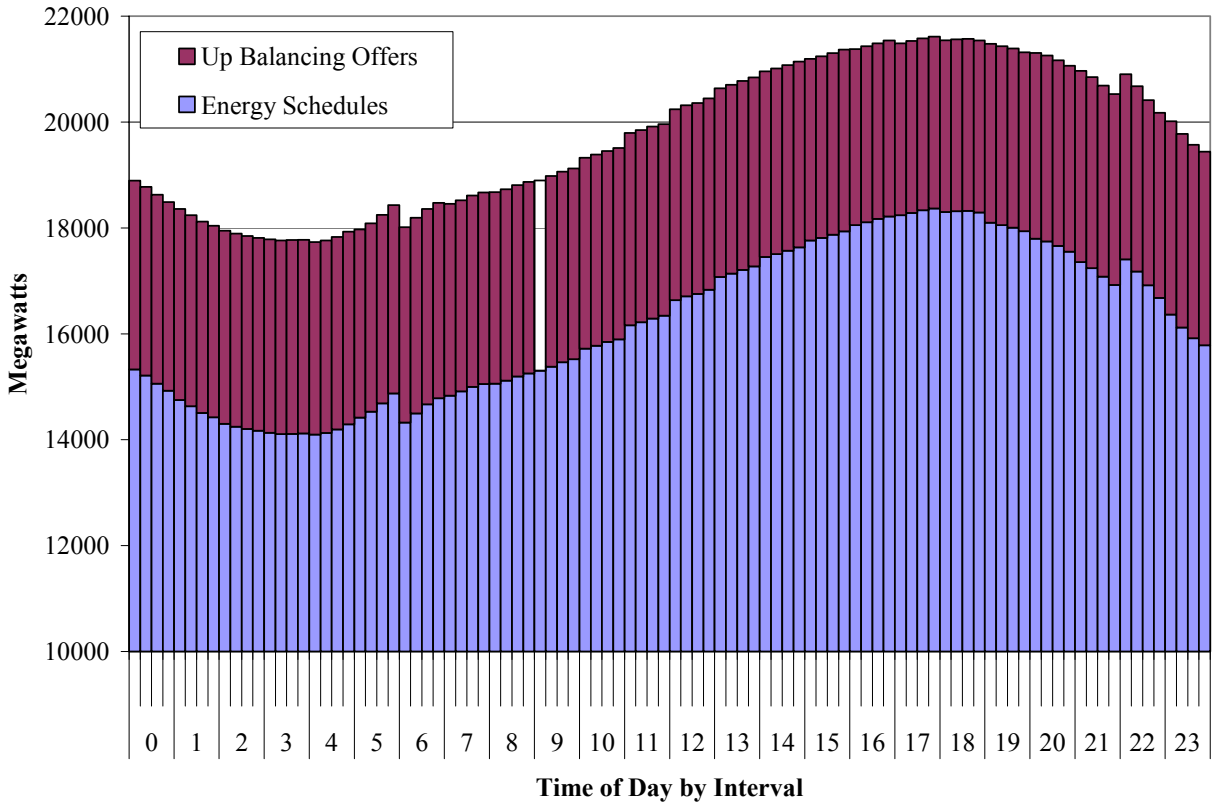


Figure 32 shows that the largest QSEs tend to schedule much more flexibly than the other QSEs. These two QSEs fully utilize the capability to schedule energy on a 15-minute basis. Like the schedules of the other QSEs, these schedules show a noticeable shift at the beginning and end of the peak bilateral contracting period from hour ending 7 to hour ending 22. However, in contrast to other QSEs, the large QSE's energy schedules show a *decrease* in the first interval of hour ending 7 and an *increase* in the first interval of hour ending 23. This demonstrates that the large QSEs take advantage of profitable arbitrage opportunities that occur at the beginning and end of the peak bilateral contracting period. Although the large QSEs do not completely counter-balance the large changes in schedules by smaller QSEs at the beginning and end of the 16 hour peak period, their scheduling patterns in those ramping hours and the fact that they generally

submit energy schedule changes on a 15-minute basis contribute to improving the performance of the balancing energy market.

Finally, the figure shows that the large suppliers tend to offer more energy (in total and as a percent of their capability) into the balancing energy market throughout the day than the smaller QSEs. This should further improve the performance of the balancing energy market by increasing its liquidity. This is evaluated further in the next subsection.

The analysis in this section shows that one of the significant issues in the current ERCOT market is the tendency of most QSEs to alter their energy schedules hourly. This tendency may be related to the fact that balancing energy bids and offers are submitted hourly and are made relative to the energy schedule. For example, if a QSE schedules 200 MW from a 300 MW resource, it may offer the remaining 100 MW in the balancing energy market. If it schedules 230 MW, it may offer 70 MW. However, if the energy schedule changes on a 15 minute basis, it may be difficult to reconcile the schedule with the hourly balancing offer, leading most QSEs to simply submit hourly schedules. This places a burden on the balancing energy market to reconcile the differences between the hourly schedules and the 15-minute actual load levels, which can result in inefficient price fluctuations.

To address this issue, we recommend changes that may increase the willingness of QSEs to submit flexible schedules (i.e., schedules that can change every 15 minutes). To that end, we have recommended that ERCOT consider introducing two scheduling options for the participants. First, it could be helpful to QSEs to allow them to submit an energy schedule for the end of the next hour that would be used by ERCOT to produce 15-minute schedule quantities by interpolating across the hour.

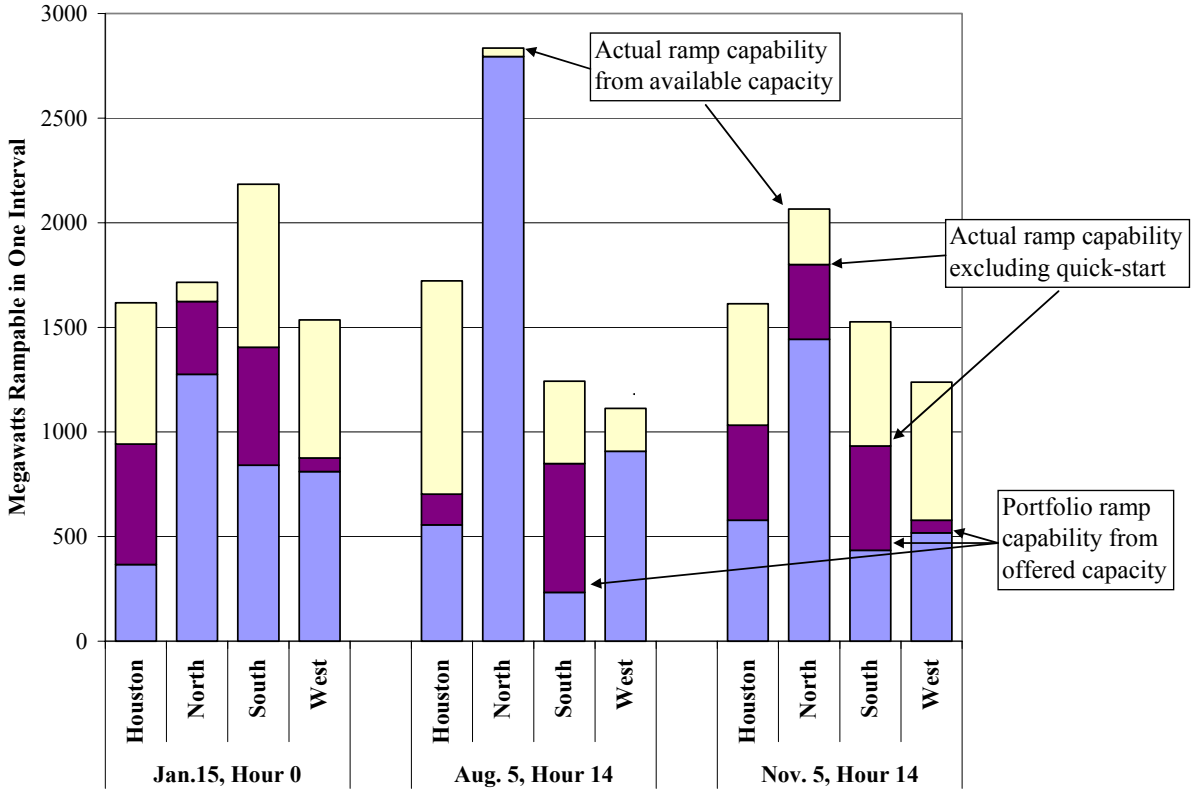
Second, we recommend that ERCOT implement an optional capability for QSEs to automatically adjust their hourly balancing energy offers for the changes in their 15-minute schedules. This adjustment would assume that intra-hour increases in energy schedules are supplied from the lowest-cost portion of the QSE's balancing energy offer. This would help ensure that the participant's portfolio energy offer is consistent with its energy schedules when the energy schedule is changing each interval. These changes would likely increase the portion of the load that is scheduled flexibly and improve the performance of the balancing energy market.

C. Portfolio Ramp Limitations

The volatility of the balancing prices in each interval is primarily related to the balancing energy deployments. However, as explained in this subsection, this volatility can be exacerbated when the portfolio ramp rates are binding. Portfolio ramp rates are constraints QSEs submit with their balancing energy offers to limit the quantity of balancing up or balancing down energy that may be deployed in one interval. These ramp rates are important because they prevent a QSE from receiving deployment instructions that it cannot meet physically. Large changes in balancing deployments from interval to interval can cause the ramp rate constraints to bind, preventing the deployment of lower-cost offers and compelling the deployment of higher-cost offers from other QSEs. Ramp rate constraints can also be limiting when resources are instructed to ramp down quickly, although this is less common.

Portfolio ramp rates are a QSE's means for representing on a portfolio basis the physical ramp capability of the units within its portfolio. In Figure 33, we compare the portfolio ramp rates to actual physical ramp capability for a single hour.

Figure 33: Portfolio Ramp Rates and Ramp Capability
Three Sample Hours in 2003



These three hours were selected because they are representative of a variety of conditions during 2003. They occur in the winter, summer, and fall. Two of the hours occur in an afternoon peak hour while one is an off-peak night-time hour. The bars in the chart show the quantity of unscheduled capacity from each zone that can be ramped in one interval. The bottom portion of the bar represents the balancing up offers that are available in one interval based on the portfolio ramp rate submitted by the QSE. The next portion of the bar is the additional energy that is physically available from online resources given their unit-level physical ramp limitations. The top portion of the bar is the additional energy that can be provided within one interval from offline quick-start resources. Hence, the total height of each bar is the amount of energy that could physically be provided within one interval (i.e., "rampable" capacity) from all available resources while the bottom portion of each bar shows the lower amount of energy available in the balancing energy market due to the tighter portfolio ramp limitations.

This analysis identifies the potential effects of the current portfolio offer structure. In Houston, 64 to 77 percent of the physically rampable capacity was not available in one interval based on portfolio offers and ramp rates. A large share of this difference may be explained by offline quick-start resources not being offered in the balancing energy market. We discuss below why this is likely the case.

In the South and West Zones, usually less than half of the physically rampable capacity is actually available to the balancing energy market in one interval. As in Houston, a large share of this difference may correspond to offline quick-start resources. The North Zone showed the highest levels of rampable portfolio offers as a percentage of the physically rampable capacity – over 70 percent. It is notable that the North Zone has less quick-start capacity as a percentage of its total capacity.

These results are not surprising given the portfolio bidding structure in the current balancing energy market. When a QSE receives balancing energy deployments from its portfolio, it will naturally prefer to increase output on its lowest-cost resources to satisfy the deployment. To the extent that a supplier's physical ramp capability is on higher-cost resources (e.g., gas turbines) that it would not prefer to dispatch before its lower-cost resources, it is not rational for the

portfolio ramp rate to include that ramp capability. The following example illustrates the problem that QSEs face in offering multiple resources using a single portfolio ramp rate.

Suppose that a QSE has a portfolio consisting of just two units, one coal-fired and one natural gas-fired. The coal unit can produce 100 MW of additional energy at a marginal cost of $16/MWh, while the natural gas unit can produce 100 MW of additional energy at a marginal cost of $50 per MWh. Assume that both units have a ramp rate of 5 MW per minute, allowing each unit to dispatch 50 MW in one interval and to be fully dispatched in two intervals. If the market clearing price in the next interval is between $16 per MWh and $50 per MWh, the balancing energy market will deploy the 100 MW of lower costs energy from the coal unit. However, to satisfy such a deployment, the generator would have to dispatch 50 MW from the natural gas unit since the coal unit can only ramp 50 MW in one interval (10 minutes times 5 MW per minute). To address this risk, the QSE may submit a portfolio ramp rate of 5 MW per minute, although this introduces potential opportunity costs when prices are higher than $50 per MWh and it is profitable to dispatch the natural gas unit. This simple example illustrates why the portfolio offer structure can generally lead suppliers to submit ramp rates that are substantially less than the maximum physical ramp rate or find other ways to address the difficulties associated with the portfolio bidding framework.

One of the significant benefits of adopting the Texas Nodal markets would be the ability of suppliers to make offers from specific units with unique ramp rates. This would provide increased flexibility and more efficient dispatch of the system. To improve the operation of the current market, we recommend that ERCOT encourage QSEs to submit multiple portfolio offers with independent ramp rates by defining “sub-QSEs”. This allows increased flexibility to offer higher-cost, fast-ramping resources without the risk that they may have to be dispatched at a loss. Additionally, this provides a means for QSEs to offer the capability from their gas turbines in the balancing energy market. Although this is currently allowed, very few QSEs take advantage of this opportunity. An additional recommendation related to gas turbines is made in the following subsection.

An additional factor that limits QSEs ability to submit portfolio ramp rates that make maximum use of their physical resources is that energy schedule changes are not currently considered when

the balancing energy market is cleared. QSEs offer balancing up and balancing down energy quantities that are constrained by its portfolio ramp limitation relative to its balancing energy deployment in the prior interval.

For instance, if a QSE has 100 MW of cleared UBES in the previous interval and 800 MW of up offers in the current interval with a ramp rate of 45 MW per minute, the QSE could sell up to a total of 550 MW (= 100 MW + 45 MW per minute * 10 minutes) in the current interval. If 550 MW is cleared in the current interval, the remaining 250 MW could be available in the subsequent interval. Thus, the ramp rate limits the change in the *deployed* quantity from one interval to the next without respect to changes in the energy schedule. In the previous example, if the QSE's energy schedule increased by 400 MW in the second interval, the QSE would actually be responsible for increasing output by 850 MW (= 400 MW schedule change + 450 MW change in cleared balancing energy). If 450 MW represents the maximum amount the QSE can ramp in one interval, the QSE would need to submit a ramp rate of 5 MW per minute so that the total change in output is physically feasible and does not exceed 450 MW (= 400 MW schedule change + 50 MW change in cleared balancing energy).

However, this strategy of simply lowering the portfolio rate can prevent the unit from receiving deployments that are fully consistent with its offer prices, reducing the supplier's profits and the efficiency of the overall market in two ways. First, this would reduce the ability of the balancing market produce balancing energy deployments that would offset prior deployments and schedule changes. In the example above, for instance, the 5 MW per minute ramp limitation would constraint the total change in balancing deployments to be less than 50 MW even though the QSE could clearly accept a reduction in its UBES deployment from the prior interval by 100 MW to zero by simply increasing its output only 300 MW rather than the 400 MW called for by the increase in its energy schedule.

Second, the portfolio offers and ramp rates are set every hour and cannot change each interval. Thus, in the three subsequent intervals of the same hour in the example above, the QSE would be limited to a total change balancing deployments of 50 MW in each interval resulting in a maximum increase in UBES over the hour of 200 MW. In reality, once the QSE increases its output to satisfy its 400 MW schedule change, it could then accept increasing UBES

deployments of as much as 450 MW in each interval (and the deployment of its entire 800 MW portfolio offer over two intervals).

To avoid these issues, the QSE could choose to submit a portfolio ramp rate that does not account for its changes in energy schedule (45 MW per minute in our example). However, it would then risk uninstructed deviation penalties in the first interval if it receives a physically infeasible balancing energy deployment. Hence, the current application of the portfolio ramp rate constraints makes it impossible for QSEs to submit an accurate ramp rate for all four intervals when its energy schedule is changing significantly at the top of the hour.

To address this issue, we recommend that ERCOT modify its SPD software for the balancing market model to account for the ramp capability that is utilized (or created) when the energy schedule increases or decreases. For example, assume that the hypothetical QSE discussed above can ramp up or down by 45 MW per minute or 450 per interval. If its energy schedule changes by 400 MW at the top of the hour, the SPD should recognize that its now has the capability to balance up in the first interval of the hour by only 50 MW (450 MW – 400 MW) and to balance down by 850 MW (450 MW + 400). The recognition that it has an increased capability to balance down by simply not increasing its output consistent with its energy schedule may sometimes provide SPD valuable additional flexibility in making balancing energy deployments.

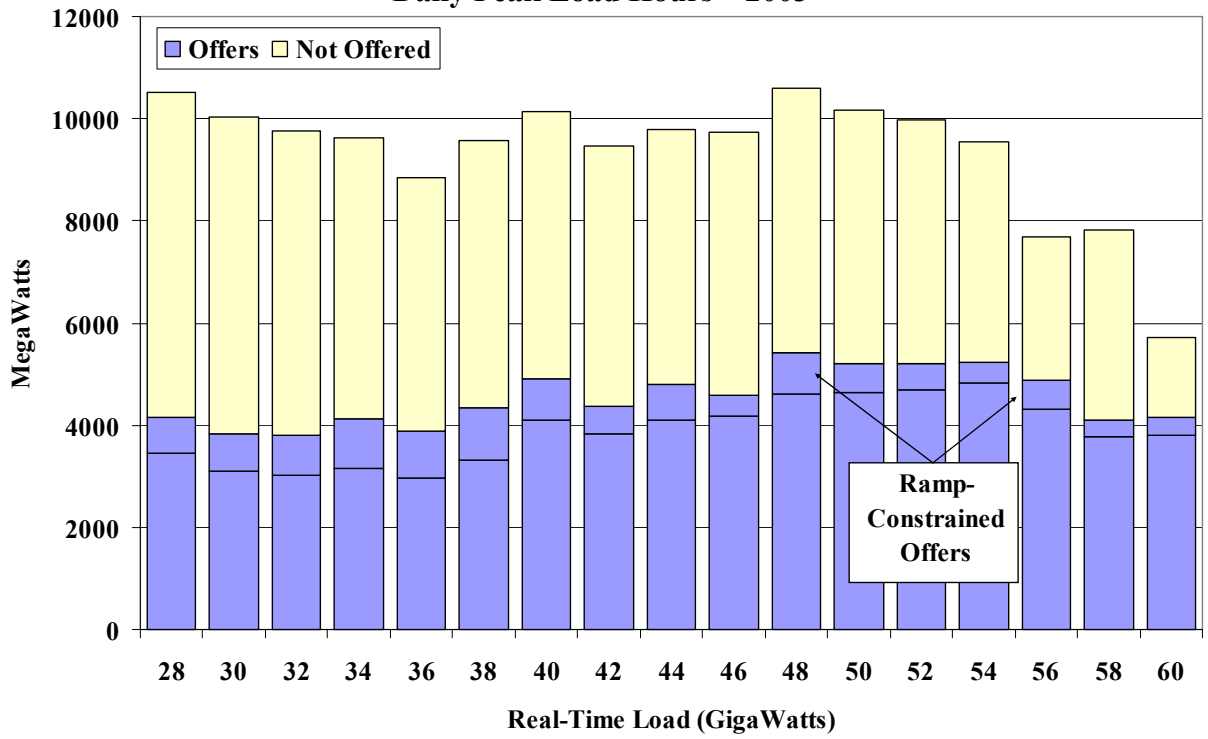
In the second interval of the hour, SPD would then recognize that the QSE can now increase or decrease its balancing energy deployments by 450 MW without the QSE having to modify its portfolio ramp limitation. Hence, this change can potentially increase the flexibility of the energy offered in the balancing market and improve the balancing energy market's performance.

D. Balancing Energy Market Offer Patterns

In this section, we evaluate balancing energy offer patterns by analyzing the rate at which capacity is offered to supply balancing energy. In Figure 34, we show the average amount of capacity offered to supply balancing up service relative to all available capacity. The offered capacity is divided into that which is ramp-constrained, and would not actually be capable of supplying balancing energy, and that which is non-ramp-constrained, and thus would be available to supply balancing energy. Capacity is considered to be available if it is either

committed or quick-start and if it is not scheduled to provide energy, reserves, or up-regulation. The data is arranged by corresponding load levels for peak hours.[13] The value shown on the horizontal axis is the upper end of the load range over which the average is computed. For example, when the value 34 GW is shown on the horizontal axis, the stacked bar includes all days when peak load is between 32 GW and 34 GW.

Figure 34: Balancing Energy Offers versus Available Energy
Daily Peak Load Hours – 2003



As the figure indicates, the average amount of available capacity at the peak hour begins to decline noticeably when peak load exceeds 54 GW. Average available capacity (whether offered or not offered in the balancing energy market) declines to less than 6 GW at peak load hours exceeding 58 GW. This available capacity is energy that can be produced in excess of the scheduled energy, reserves, and up-regulation for a QSE. Hence, a substantial surplus remains even under peak conditions based on the online resources and available quick-start resources.

13 More precisely, available capacity and balancing-up offers were ascertained for the peak hour of each day during 2003. This data was then separated by load level and the average capacity and average balancing-up offers were calculated.

On days when peak load is between 26 GW and 54 GW, quantities offered into the balancing energy market at the peak hour rise steadily from 4 GW to over 5 GW. On days when peak load is above 54 GW, offer quantities decline as capacity is in shorter supply. In percentage terms, the portion of available capability actually offered at the peak hour increases from 44 percent on days with peaks below 36 GW to 73 percent on days when peak load is greater than 58 GW.

Figure 34 shows the average share of the offer quantity that can be ramped up in one interval, indicating that participants generally offer slightly more than this level. On average, only 300 MW to 800 MW of offers were unavailable in a single interval due to ramp constraints. The fraction of offers unavailable due to ramp constraints was smaller at higher actual load levels.

The results shown in Figure 34 indicate that there is a substantial amount of available balancing energy that is not offered, which increases the volatility of balancing energy prices in intervals when balancing deployments are relatively large. There could be a number of reasons for these results. First, the issues related to ramp rates discussed in the prior subsection can affect the offer levels. To the extent that a supplier's portfolio includes slower-ramping low-cost resources, the supplier may not offer a significant share of its higher-cost resources. The supplier faces the risk that it will receive a balancing energy deployment that exceeds the ramp capability of the low-cost resources, compelling it to dispatch its high-cost resources at a loss.

In addition to this general issue related to the portfolio bidding framework, the SPD software had a specific issue that may have caused QSEs to offer only a portion of their available energy. In 2003, the SPD software assumed that a supplier could ramp back down to its schedule immediately (i.e., the recall ramp rate was not respected).[14] For example, if a portfolio deployment of 300 MW was made over 3 intervals, SPD may instruct the supplier to return to its schedule in the fourth interval. If the supplier can only ramp down 100 MW per interval, it will show an uninstructed deviation for two intervals. By offering what could be ramped in one SPD interval, a participant will minimize its risk of uninstructed deviations caused by infeasible recall instructions. This strategy is effective because it will ensure that the supplier's deployment will not be significantly larger than the amount that can be recalled in one interval.

[14] On June 2, 2004, an upgrade was made to ERCOT's balancing energy market software to enforce recall ramp rates, which should address this issue.

Second, it is very difficult to offer gas turbines in the balancing energy market effectively. The available capability in Figure 34 includes quick-start offline resources, primarily comprised of gas turbines. If these quantities were eliminated from the figure, it would show that a much higher portion of the available balancing energy was offered. The current balancing energy market rules present significant challenges for owners of gas turbines due to timing and minimum run-time considerations. With regard to timing, the current balancing energy market rules do not provide adequate advance notice for some suppliers to reliably start gas turbines in response to the balancing energy market instructions. In addition, there is no assurance that prices in subsequent intervals will support the continued operation of the gas turbine.

As shown above, balancing energy prices frequently spike in the first or last interval of an hour, before decreasing significantly. This could cause a supplier with a gas turbine that is satisfying its portfolio instruction to have to turn on the gas turbine for the one interval, and then keep it on for the rest of its minimum run time at a loss before it may shut down. Hence, it is understandable that some suppliers would not offer the energy that may be available from their gas turbines in the balancing energy market (or only offer it under higher load conditions when balancing energy prices could be expected to be sustained for multiple intervals).

Third, the lack of offers could reflect withholding by ERCOT participants. Although this is possible, the fact that the percentage of available capability offered into the market is the lowest in the lowest load periods does not support this hypothesis. The incentive to withhold should be highest under peak demand conditions when the withholding would have the largest effect on prices. In fact, the percentage of available capability offered into the market steadily rises with load and is at its highest on days with the greatest demand. Nevertheless, these offer patterns warrant further analysis and investigation.

To provide additional insight regarding the possibility that the offer patterns may raise competitive concerns, we next examine whether large and small suppliers act in systematically different ways in terms of the available capacity they offer into the balancing energy market. If large suppliers offer less of their available capacity, particularly under peak conditions, that could be an indication of market power. Figure 35 shows the balancing up capability relative to the balancing up offers divided between large suppliers and small suppliers. The large suppliers

category includes QSEs associated with the largest two owners of generating capacity in ERCOT, whereas all other QSEs are included in the "small" suppliers category.

We emphasize that this analysis will not support a definitive finding regarding potential competitive issues related to these offer patterns. Depending on load and contractual obligations or location, some small suppliers may have more market power than some of the large suppliers. However, the analysis does provide useful information to consider in conjunction with the other results in this report to determine whether additional investigation is warranted.

Figure 35: Balancing Energy Offers versus Available Energy in 2003
Large and Small Suppliers -- Daily Peak Load Hours

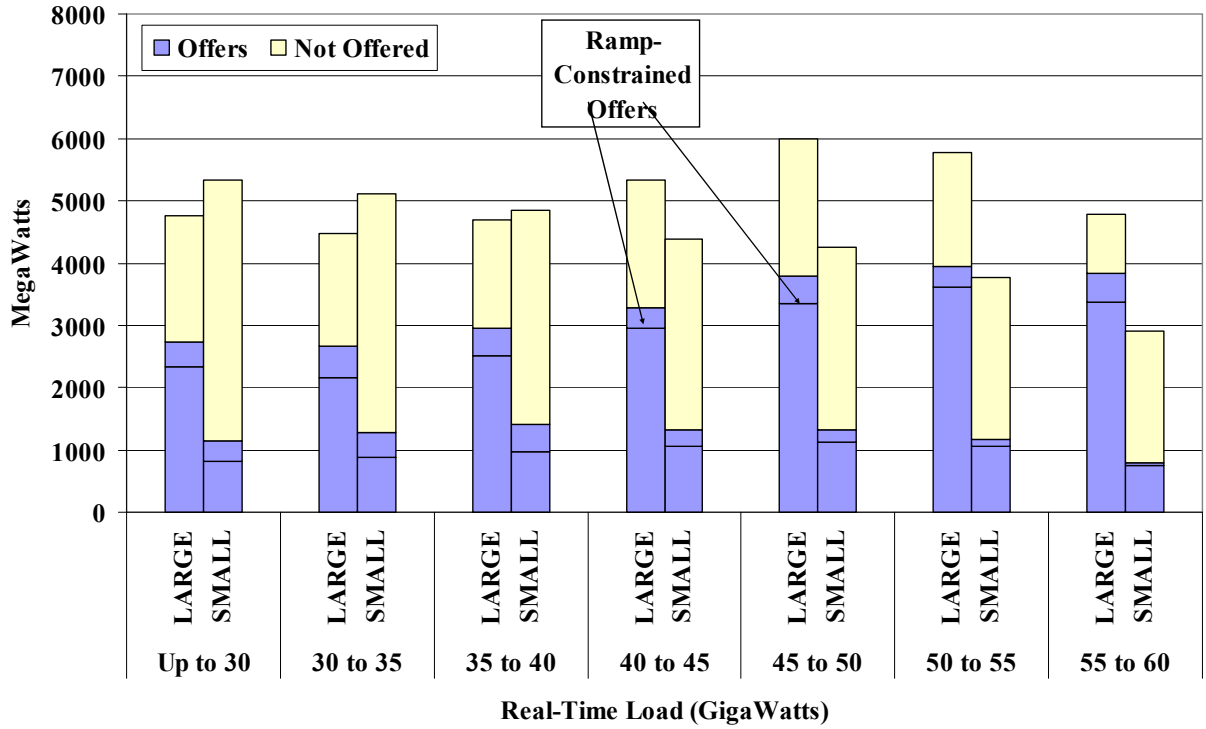


Figure 35 shows that available balancing energy for small suppliers tends to decrease slightly with actual load, whereas available balancing energy for large suppliers actually rose slightly to reach its highest levels when real-time load was between 45 GW and 55 GW. This indicates that the commitment of online resources by large suppliers tends to increase more rapidly than the real-time load when the load rises into these ranges.

The amount offered into the balancing energy market by large participants trends upward while the amount offered by small suppliers is flat to slightly downward sloping. At the lowest load

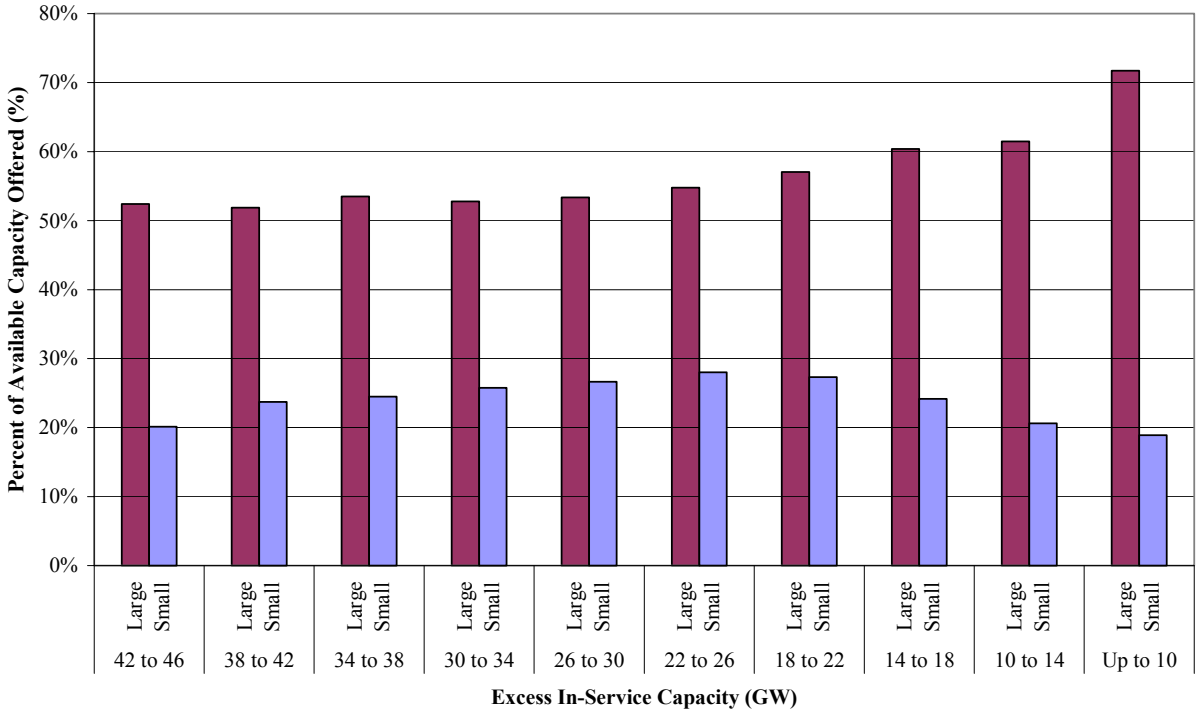
levels, large participants offered an average of 2700 MW or 57 percent of their available capacity. At the highest load levels, large participants offered an average of 80 percent of their available capacity. At all load levels, small participants offered an average of less than 32 percent of their available capacity. Both large and small suppliers offered only slightly more balancing energy than the portfolio ramp constraints allow to be deployed in a single interval.

Because large participants are most likely to have market power and would generally have the largest incentive to withhold during periods of high demand, the fact that large participants offer their generation at a higher rate than small suppliers during the highest demand periods does not raise competitive concerns. If any concerns are raised by this analysis, the concerns would pertain to the conduct of the smaller suppliers. Investigating the specific conduct of these smaller suppliers is beyond the scope of this report.

We performed one final analysis of these offer patterns in which we calculate the percentage of available energy offered in the balancing energy market by large and small suppliers under different excess capacity conditions. The prior two figures examine offer patterns relative to actual load levels. However, volatile prices frequently occur in the spring and fall seasons at lower load levels. Tight conditions can occur during the shoulder months when a relatively large quantity of generation is out of service for planned maintenance. Like the peak conditions during the summer, tight conditions at other times of the year should be characterized by relatively low excess capacity levels.

Figure 36 shows the results of this analysis for the largest four QSEs and all other smaller QSEs. The excess capacity metric shown in this figure is equal to the total available in-service capacity as shown in the day-ahead resource plans less the actual load and ancillary services requirements. These results confirm the prior results, indicating that the largest suppliers tend to offer a higher share of their available energy in the balancing energy market than smaller suppliers. Additionally, this ratio increases as the quantity of excess capacity in the market decreases (i.e., as system conditions become tighter). These results do not suggest broad competitive concerns regarding the conduct of the largest suppliers, although the portion of the available energy offered under most conditions is disappointing at close to 50 percent.

Figure 36: Ratio of Balancing Energy Offers to Available Energy
Large and Small Suppliers -- 2003



The ratio of available energy offered by smaller suppliers remains between 20 and 30 percent under most excess capacity levels, falling below 20 percent under the tightest market conditions. Investigation of the causes for these low offer levels, particularly among the smaller suppliers, is warranted due to the effect these offer patterns have on the performance of the balancing market and the possibility that they may include discrete instances of withholding.

In conclusion, we believe that current market design does not provide an efficient means for gas turbines to be dispatched and priced for the reasons described above. In addition, we conclude that the hourly portfolio offer provisions limit the availability of output that cannot be ramped in a single SPD interval. To address both of these concerns, we recommend that ERCOT encourage QSEs to submit multiple, independent portfolio offers with different ramp rates by defining sub-QSEs. This allows a supplier to submit an offer for the lower-cost, slow-ramping portion of its resources independent of its higher-cost resources, including its gas turbines. Hence, suppliers can use this capability to more freely make each segment of their resources available in the balancing energy market at prices that are compensatory and ramp rates that are achievable.

E. Analysis of Resource Plans

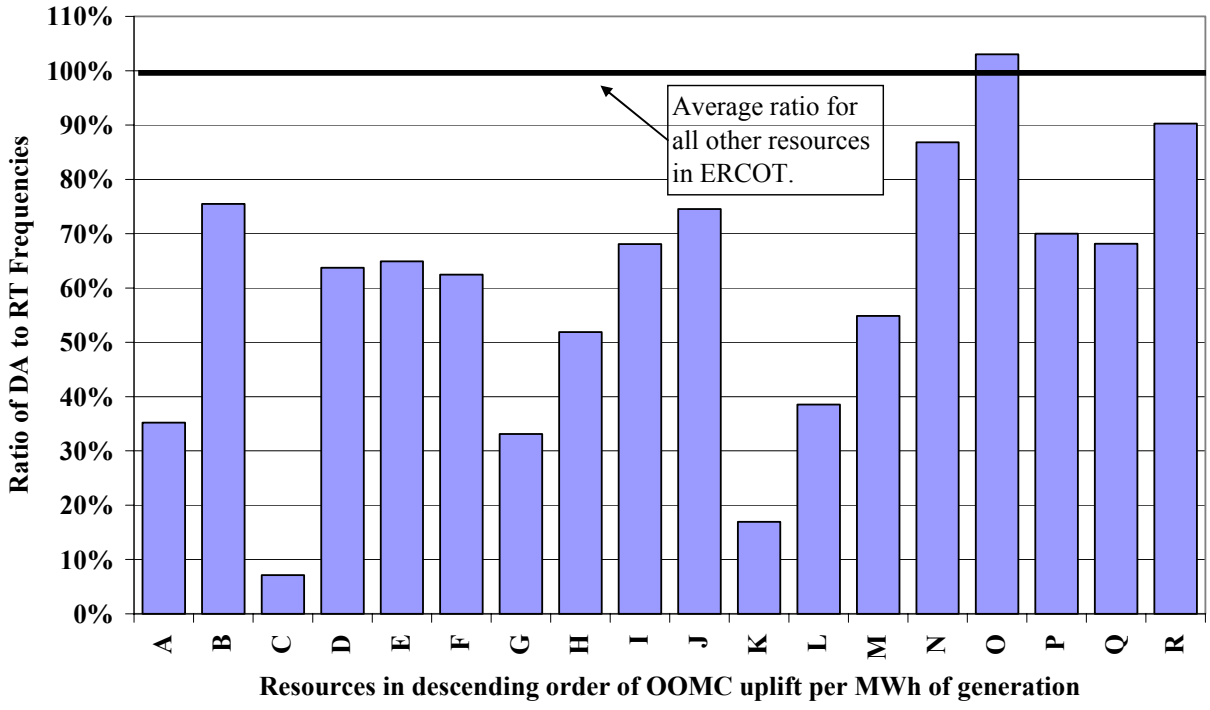
The submission and use of resource plans raise general concerns because resource plans are not financially binding, yet they are used by ERCOT to take commitment actions that can have significant cost implications. Hence, if a market participant can affect ERCOT's actions and the revenue it receives by submitting resource plans that do not represent efficient generator commitment and dispatch, it may do so at no cost since the plans are not binding. In this subsection, we analyze participant resource plans to evaluate whether the market protocols may provide incentives for strategic conduct. In particular, we evaluate units that are frequently committed out-of-merit or frequently dispatched out-of-merit. Such units receive additional payments from ERCOT and participants may engage in strategies to increase these payments.

We first analyze the behavior of suppliers that are the primary recipients of payments by ERCOT for out-of-merit capacity. Out-of-merit commitment ("OOMC") occurs when ERCOT instructs a unit that is not committed in the QSE's day-ahead resource plan to start in order to ensure sufficient capacity is available in real time to meet the forecasted load and manage transmission constraints. When suppliers provide OOMC, they receive payments from ERCOT that correspond to an estimate of the cost of starting the unit plus an amount to contribute to the estimated costs of running at some minimal level. To the extent that the unit provides output in real-time, the balancing energy revenue from these sales is retained by the supplier. Therefore, if a unit is frequently needed to for OOMC, a supplier has the financial incentive to refrain from showing the unit as committed in the day-ahead resource plan to compel ERCOT to commit the unit. This supplier can do this because it can always commit the units itself before real time.

In other words, if the supplier is anticipating that its unit will be profitable in the real-time market (in the sense of earning at least enough to cover start-up and running costs), the supplier may be better off not showing the unit as committed in the day-ahead resource plan. By not committing day-ahead, the participant can wait for the OOMC instruction, receive the OOMC payment, and still earn the energy-market revenues in real-time. If it does not receive the OOMC instruction, it does not forgo any profits in the real-time energy market because it can commit itself in a subsequent resource plan before real-time. This is generally a risk-free method to attempt to receive additional revenue through OOMC for most units, excluding those requiring a relatively long period of time to start-up.

Because of these incentives, we would expect suppliers that anticipate having units committed out-of-merit to avoid showing the units as committed until after the out-of-merit commitments are announced. To test this hypothesis, we examined the patterns of commitment for units that receive substantial OOMC payments. Figure 37 shows the ratio of day-ahead resource plan commitments to actual real-time commitment during 2003 for the 18 resources receiving the largest OOMC payments per MWh of production.[15] Hours when the resources are under OOMC or OOME instruction are not included in order to assess systematic changes made voluntarily by market participants. The units are shown in decreasing order of payments received on a per MWh basis—from $82 per MWh for the units on the far left to $11 per MWh for the units on the far right. To show how the commitment of these units compares to all other units in ERCOT, the figure also shows the capacity-weighted average ratio of day-ahead to real-time resource plan commitments for all units.

Figure 37: Ratio of Day-Ahead to Real-Time Resource Plan Commitments*
Frequent OOMC Resources – 2003



* Excluding hours where resources are under OOMC instructions or dispatched out-of-merit. The resources shown in this figure are listed in Appendix A.

15 We exclude resources that received payments that total less than $10 per kW-year of capacity or averaged less than $10 per MWh of generation.

Of the 18 resources shown in Figure 37, 17 have ratios of less than 100 percent, ranging from 8 percent to 90 percent. One resource had a ratio of 103 percent. These results are consistent with the hypothesis described above. In contrast, the average ratio for all other units is 99 percent, reflecting a much higher consistency between the day-ahead and real-time resource plans.

For the resources shown in Figure 37, uplift payments for OOMC are substantial enough to provide significant incentives to behave in ways that maximize the likelihood of receiving them. Figure 37 suggests that QSEs with resources that frequently receive OOMC instructions regularly delay the decision to commit those units until after ERCOT determines which resources to select for OOMC. This pattern has several deleterious effects on the market. First, ERCOT frequently incurs OOMC costs to commit resources that are otherwise economic and that should be committed voluntarily without supplemental payments. Second, when resources are committed out-of-merit, some other resources committed in day-ahead resource plans will no longer be economic. This can result in over-commitment of the system, which is inefficient and can distort balancing energy price signals, although the QSE generally has the opportunity to modify its other commitments after it receives the OOMC instructions. Third, this conduct tends to obscure the information that ERCOT relies on to manage reliability. Ultimately, this can cause ERCOT to take a variety of actions, including making out-of-merit commitments that should not be necessary.

In our next analysis, we evaluate incentive issues associated with out-of-merit dispatch in real-time. In order to resolve intrazonal congestion in real-time, ERCOT will increase or decrease a unit's output (out-of-merit energy or "OOME") to reduce the flow on a constrained transmission facility within a zone. When the unit is dispatched up in this manner (i.e., OOME up), it receives payments corresponding to the higher of the estimated running cost of the out-of-merit portion of the unit (plus a margin), or the balancing energy price. Although the potential profits are limited by the formula used to calculate the OOME payment, the system can still provide incentives to schedule resources strategically.

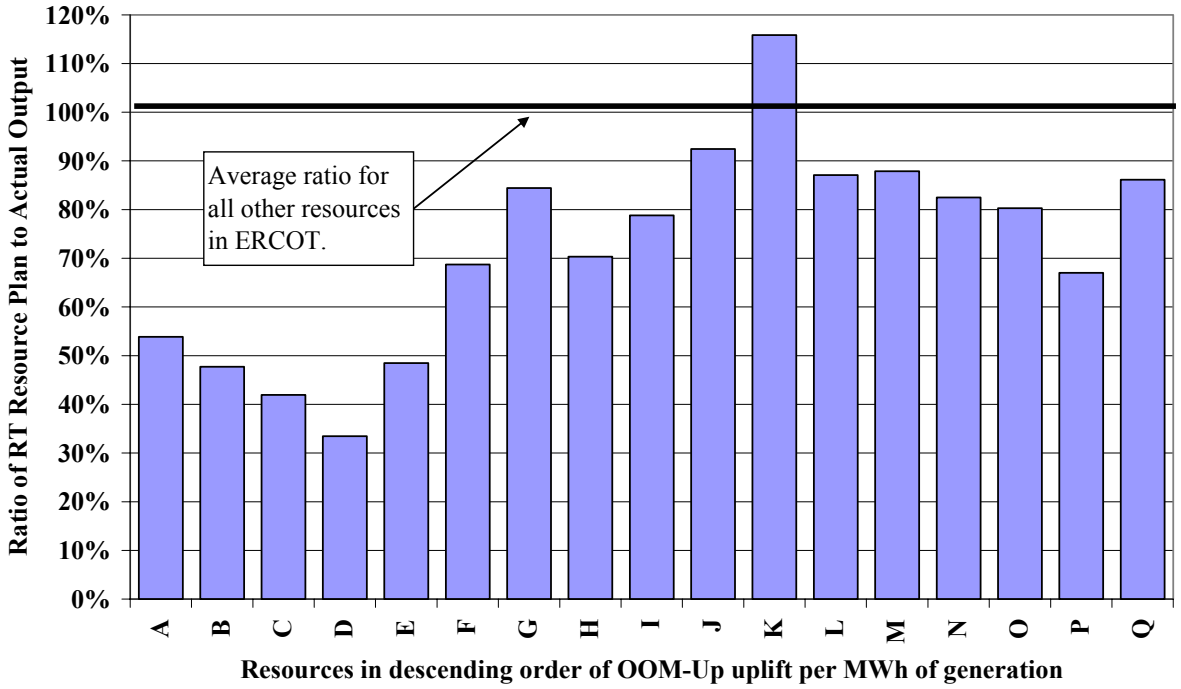
If a supplier is able to predict which of its units may be dispatched out-of-merit, it may under-schedule those units and over-schedule other units in its portfolio.[16] Although this resource plan

16 "Scheduling" in this context refers to the unit-level projected output levels in the QSEs' resource plans.

output may not be efficient, it can be effective at compelling an OOME instruction and the associated uplift payment. Following the OOME instruction, the supplier can adjust its over-scheduled units to restore an economic dispatch pattern. If the supplier can accurately predict when the units will be called out-of-merit, this strategy can generate significant uplift payments. When the unit is not called for out of merit dispatch, the supplier can adjust the output levels of the units in its portfolio to correct the inefficient schedule.

Under this type of strategy, one would expect that units often needed to resolve congestion would be frequently under-scheduled. To test for this strategy, Figure 38 shows the ratio of real-time scheduled output for the unit in the resource plan to actual generation for the 17 units that received the highest average payments (per MWh) for OOME up in 2003.[17] To include only the scheduling and dispatch decisions made solely by the supplier, the ratio does not include hours when the resource was under OOMC or OOME instructions.

Figure 38: Ratio of Real Time Resource Plan Output to Actual Generation
Frequent OOME up Resources – 2003



* Excluding hours where resources are under OOMC or OOME instructions. The resources shown in this figure are listed in Appendix A. Real-time resource plan output is the projected output levels in the real-time resource plans.

17 To focus on the most significant units, the analysis excludes resources where total uplift received was less than $2 per kW-year of capacity or the average was less than $2 per MWh of generation.

The 17 resources shown in Figure 38 are presented in decreasing order of average payments, from $19 per MWh for the unit on the far left to $2.72 per MWh for the unit on the far right. The generation-weighted average ratio of real-time resource plan output to actual generation for the whole ERCOT market is also shown for reference.

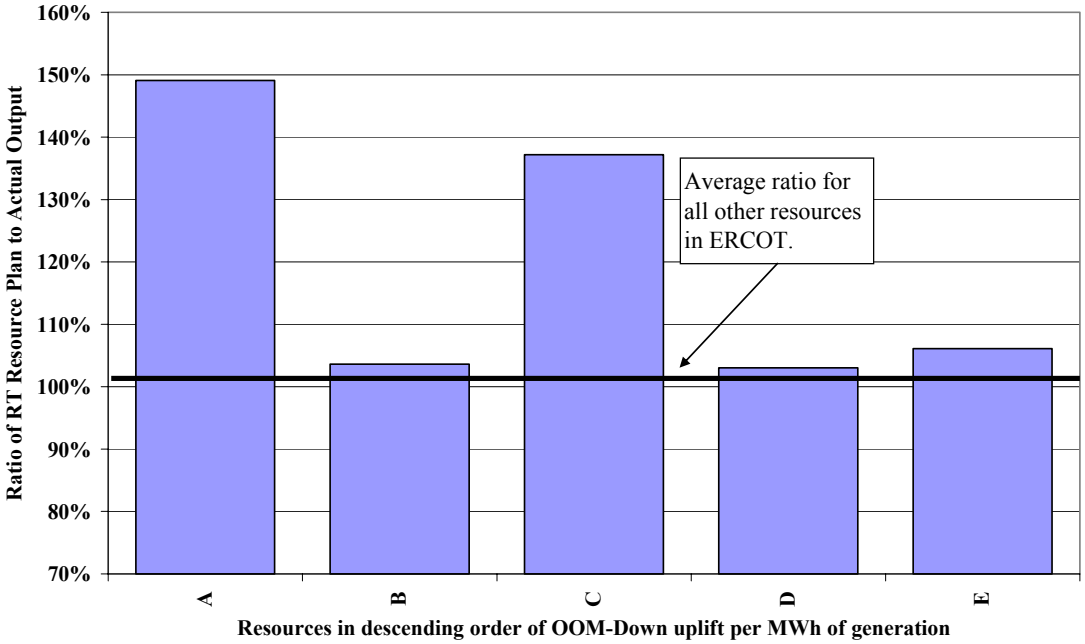
Of the 17 resources shown in Figure 38, 16 have ratios of less than 100 percent, ranging from 33 percent to 92 percent. Thus, all but one of them operate at higher output levels in real time than they were scheduled to run in the real-time resource plan. The other units in ERCOT had a ratio of 101 percent in 2003, reflecting, on average, consistency between the scheduled output and actual generation. The data suggests that resources frequently providing OOME up are regularly included by the QSEs in the real-time resource plans at output levels that are significantly lower than their actual output. This is consistent with the hypothesis that the OOME procedures may provide inefficient incentives that lead QSEs to submit inaccurate resource plans.

We next evaluate the incentives associated with providing OOME down. The incentives associated with rules for OOME down payments are the reverse of the incentives for OOME up payments. Since ERCOT pays units to reduce output from the real-time resource plan output levels, a supplier able to foresee the need for an OOME down instruction can over-schedule the unit to compel the OOME down action by ERCOT. If the OOME down settlement rules provide strong incentives to engage in this conduct, the units that frequently receive OOME down instructions should be consistently over-scheduled. However, we would note before presenting our analysis that the magnitude of payments for OOME down is far lower than the magnitude of uplift payments for OOME up.

Figure 39 shows the ratio of real-time resource plan output to actual generation for five select resources that earned the highest average payments for providing OOME down (on per MWh basis) in 2003.[18]

[18] This analysis excludes resources with uplift payments totaling less than $1 per kW-year of capacity or averaging less than $1 per MWh of generation. This analysis also excludes cogeneration and renewable resources.

Figure 39: Ratio of Real-Time Resource Plan Output to Actual Generation
Frequent OOME down Resources – 2003



* Excluding hours when resources are under OOMC or OOME instructions. The resources shown in this figure are listed in Appendix A. Real-time resource plan output are the projected output levels in the real-time resource plans.

Figure 39 shows only five units because the OOME down payments for the remaining units were *de minimus*. The five resources are shown in decreasing order of the average OOME down payments received per MWh of output, ranging from $2.31 per MWh on the far left to $1.16 per MWh on the far right. For comparison purposes, the figure also shows the generation-weighted average ratio of real-time resource plan output to actual generation for all other units.

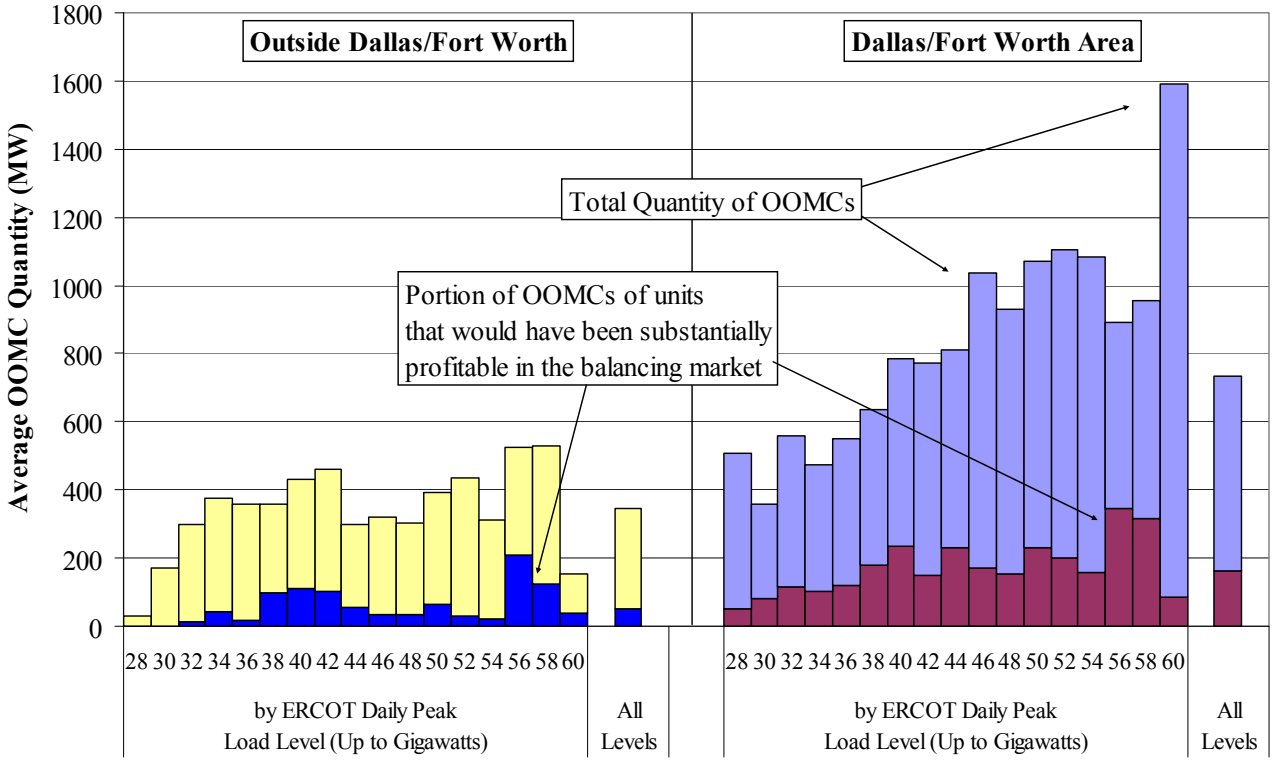
All five of the resources shown in Figure 39 have ratios of above 100 percent, ranging from 103 percent to 149 percent. However, only two of five have ratios that exceed 100 percent by a significant margin. In other words, in hours when the units are receiving no OOME down instructions or payments, two of the units are regularly scheduled at higher output levels in the real-time resource plan than the units' actual output. This is in contrast to the average ratio exhibited by other units in ERCOT of 101 percent in 2003, which reflects consistency between resource plan output and actual generation.

Four of the five resources shown in Figure 39 are located in the North-East Zone, which was created at the start of 2004. The zone was created to address the problem that resources in this

region were frequently dispatched down out of merit to manage intrazonal congestion in the North Zone. This new zone should reduce the need for OOME down and should provide better price signals to govern the commitment and dispatch of resources in the zone.

Finally, we conducted a further analysis of the local congestion and out-of-merit patterns in the Dallas/Ft. Worth area. The transmission constraints into the Dallas/Ft. Worth area are the most significant local constraints in ERCOT by most measures. Figure 40 shows two panels, one for Dallas/Ft. Worth and one for all other areas in ERCOT. Each panel shows the average quantity of OOMC relative to the peak demand levels. The figure also reports the portion of OOMC that would have been substantially profitable to self-commit based on estimated start-up costs, minimum generation costs, incremental costs, and minimum run times.[19]

Figure 40: OOMC Supplied vs. ERCOT Load Level
Dallas/Fort Worth and Other Areas, 2003



19 Profits are considered to be substantial if they would exceed the estimated minimum commitment costs of the unit by a margin of at least 50 percent. Continuous Emissions Monitoring (CEMS) data, collected by the Environmental Protection Agency, is used to estimate incremental heat rates and heat input at minimum generation levels. We also assume $4 per MWh variable operating and maintenance expenses. Whenever CEMS data is unavailable, minimum generation and incremental costs are estimated from a sample of balancing energy prices that coincide with each resource's production over the previous 90 days.

This figure shows that on average ERCOT commits a significantly larger amount of out-of-merit resources in Dallas/Ft. Worth (733 MW) than all other areas (345 MW). The figure also shows that as the demand in Dallas/Ft. Worth rises, operators must take more out-of-merit actions to maintain reliability. Both the total quantity of resources providing OOMC, as well as the portion of OOMC that was profitable to self-commit, more than doubled in the highest load hours (>50 GW) versus the lowest load hours (< 36 GW). In contrast, there is no clear relationship between OOMC quantities and demand levels outside Dallas Ft. Worth.

Approximately 20 percent of resources receiving OOMC instructions would clearly have been economic for the QSEs to self-commit. Our previous analysis of resource plan changes between the day-ahead and real-time indicates that units frequently committed out of merit are often voluntarily committed when ERCOT does not provide an OOMC instruction.

Taken together, these analyses indicate that the current procedures for OOME and OOMC provide incentives for participants to submit resource plans that do not reflect anticipated real-time operations. One change we recommend to the current markets that would mitigate these issues would be to create a zone for Dallas/Ft. Worth. This would allow a large share of the congestion that is currently managed with OOME processes to be priced more efficiently and transparently. It would also provide superior economic signals to guide investment in generation and transmission in that area. Lastly, if ERCOT were to move to a nodal market design, creating this zone would ease the transition to nodal markets where all congestion would be reflected in locational clearing prices.

We understand, however, that there would be significant issues to consider in forming such a zone, including the effect on current bilateral contracts, the need for measures to effectively mitigate market power in the area, and the equity implications of such a change. In addition, the benefits described above assume that CSCs between Dallas-Fort Worth and adjacent areas could be defined that include the key transmission constraints that currently result in OOME and OOMC actions by ERCOT. This would need to be analyzed and validated.

The most comprehensive solution for all of these issues is to implement nodal electricity markets since properly structured nodal markets would virtually eliminate the need to commit and dispatch resources out of merit. Such markets would substantially improve the efficiency of the

management of local congestion, as well as the management of interzonal congestion as discussed in detail in Section IV below. Hence, we strongly encourage the continued development and adoption of the Texas Nodal markets that are currently under consideration.

III. DEMAND AND RESOURCE ADEQUACY

The prior sections of this report reviewed the market outcomes and provided analyses of a variety of factors that have influenced the market outcomes. This section reviews and analyzes the load patterns during 2003 and the existing generating capacity available to satisfy the load and operating reserve requirements.

A. ERCOT Loads in 2003

There are two important dimensions of load that should be evaluated separately. First, the changes in overall load levels from year to year can be shown by tracking the changes in average load levels. This metric will tend to capture changes in load over a large portion of the hours during the year. Second, it is important to separately evaluate the changes in the load during the highest-demand hours of the year. Significant changes in these peak demand levels are very important because they determine the probability and frequency of shortage conditions. More broadly, the peak demand levels and capability of the transmission network are the primary factors that determine whether the existing generating resources are adequate to maintain reliability. Hence, both of these dimensions of load during 2003 are examined in this subsection and summarized in Figure 41. This figure shows peak and average loads in each of the four ERCOT zones for 2002 and 2003.

Figure 41 indicates that in each zone, as in most electrical systems, peak demand significantly exceeds average demand. The North Zone is the largest zone (about 40 percent of the total ERCOT load); the South and Houston Zones are comparable (with about 25 percent each) while the West Zone is the smallest (with about 7 percent of the total ERCOT load).

Changes in peak load levels from 2002 to 2003 are primarily attributable to weather in 2003. Temperatures reached 100° on six days in Dallas in July 2003, and in August 2003 there were eleven days when temperatures reached 100°, including 109° on August 6 and 105° on August 7. In 2002, temperatures in Dallas reached 100° during only one day in July, and never exceeded 99° in August. Houston reached 102° on August 7, 2003, compared to a peak of 97° in summer 2002. In the West Zone, temperatures reached the same peak in 2003 as in 2002: 106°. However, there were twenty-one 100° days in 2003, compared to fourteen in 2002.

Figure 41: Annual Load Statistics by Zone
2002 v. 2003

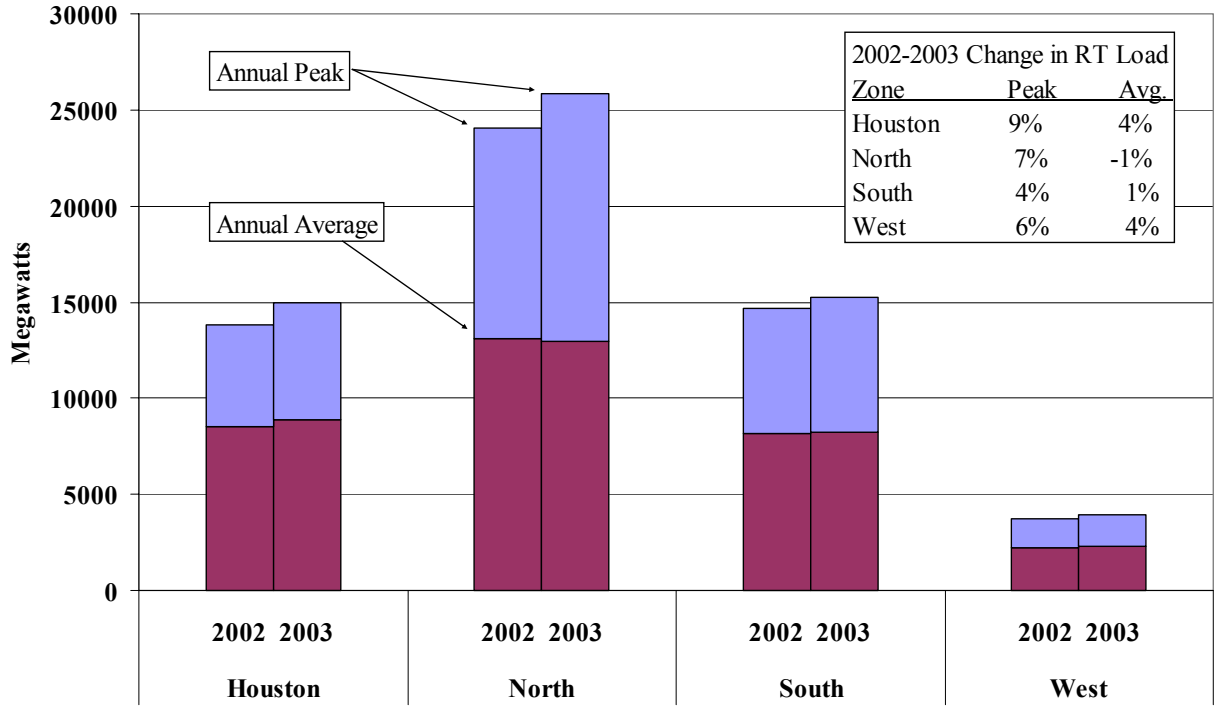


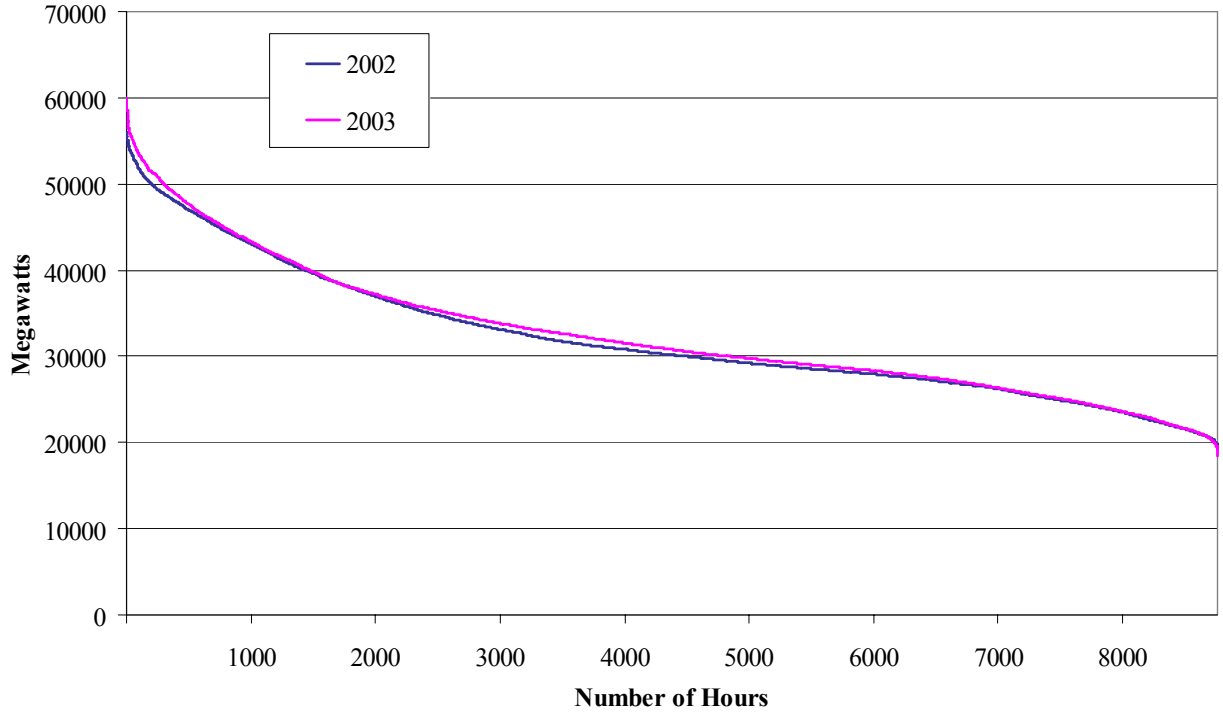
Figure 41 also shows that average loads in each zone were comparable between 2002 and 2003. This was due in part to the fact that the temperatures, with the exception of the hottest days, were relatively moderate in 2003. Cooling degrees days, on a population-weighted basis, were about 2.5 percent higher in 2002 than 2003.

Due to the high summer peak of 2003, the average load factor across the state (defined as the ratio of average demand to peak demand) declined from 57 percent to 54 percent. Similar declines were seen in each zone, with the smallest decline in the West. The highest load factors were in Houston (59.2 percent in 2003) and the West (58.8 percent). Houston has a higher load factor because the Gulf of Mexico moderates peak temperatures and the city's large manufacturing base provides a larger proportion of non-weather related demand.

To provide a more detailed analysis of load at the hourly level, Figure 42 compares load duration curves for 2002 and 2003. A load duration curve shows the number of hours (shown on the x-axis) that load exceeds a particular level (shown on the y-axis). ERCOT has a fairly smooth load duration curve, typical of most electricity markets, as most hours exhibit low to moderate

electricity demand, with peak demand usually occurring during the afternoon and early evening hours of days with exceptionally high temperatures. In 2003, the highest load hours occurred in the summer months, particularly in August.

Figure 42: ERCOT Load Duration Curve
All Hours – 2002 & 2003



As Figure 42 shows, the load duration curve for 2003 lies slightly above the curve for 2002. Indeed, overall demand increased approximately 1.4 percent in 2003 versus 2002. A significant portion of this increase is accounted for in the highest-load hours. In both years, demand at the lowest-load hours was comparable. At moderate load levels, the two years are very similar, although a modest increase in average load in 2003 is evident at most points in this range.

To better evaluate the differences in the highest-demand periods between the two years, Figure 43 shows the load duration curve for the top 5 percent of hours with the highest loads. This figure shows that differences in the peak demands between 2002 and 2003 were significantly larger than the differences in average demand.

Figure 43: ERCOT Load Duration Curve
Top Five Percent of Hours – 2002 & 2003

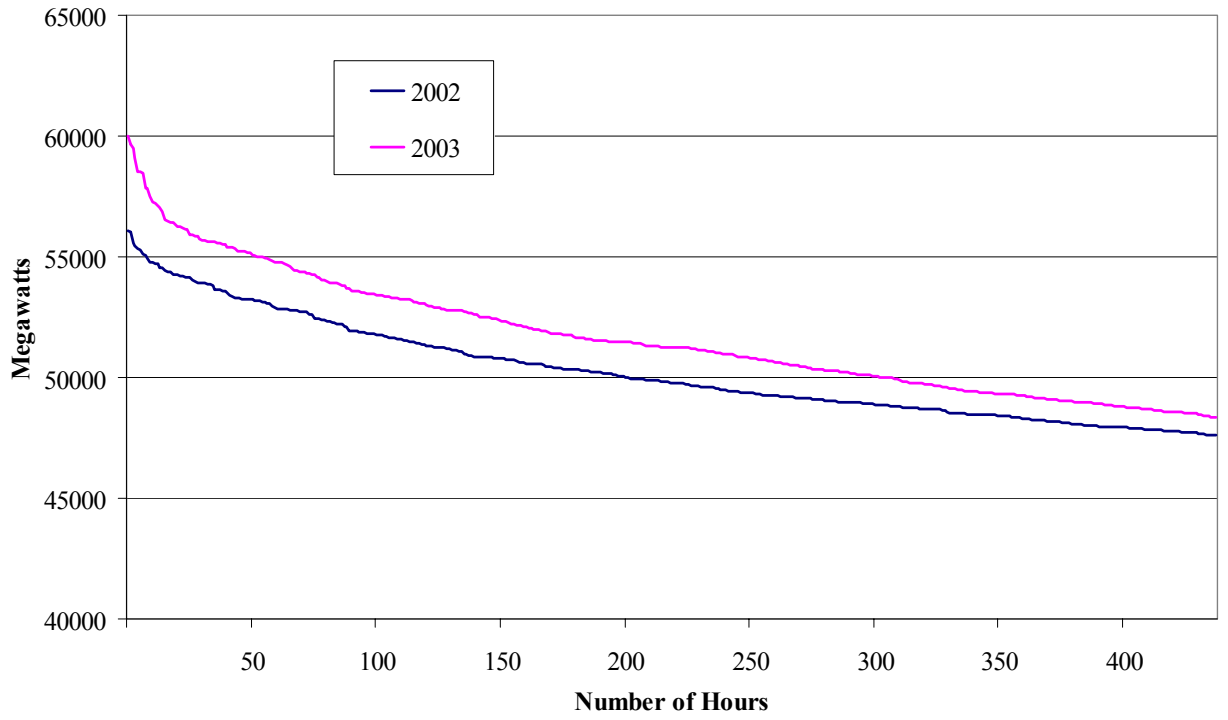


Figure 43 shows that demand exceeded 55 GW in 52 hours during 2003 and peaked very close to 60 GW. In 2002, demand was higher than 55 GW in only 7 hours and the peak demand was 56,248 MW. The same pattern prevailed at lower load levels. Demand exceeded 50 GW in 306 hours in 2003, compared to only 201 hours in 2002. These differences can largely be attributed to the increased frequency of very hot summer days in 2003. Although peak demand conditions were more frequent and more severe in 2003, they did not tend to cause sharp increases in electricity prices because the ERCOT market continues to enjoy substantial excess capacity. Hence, the peak demand conditions did not result in any shortage conditions in ERCOT.

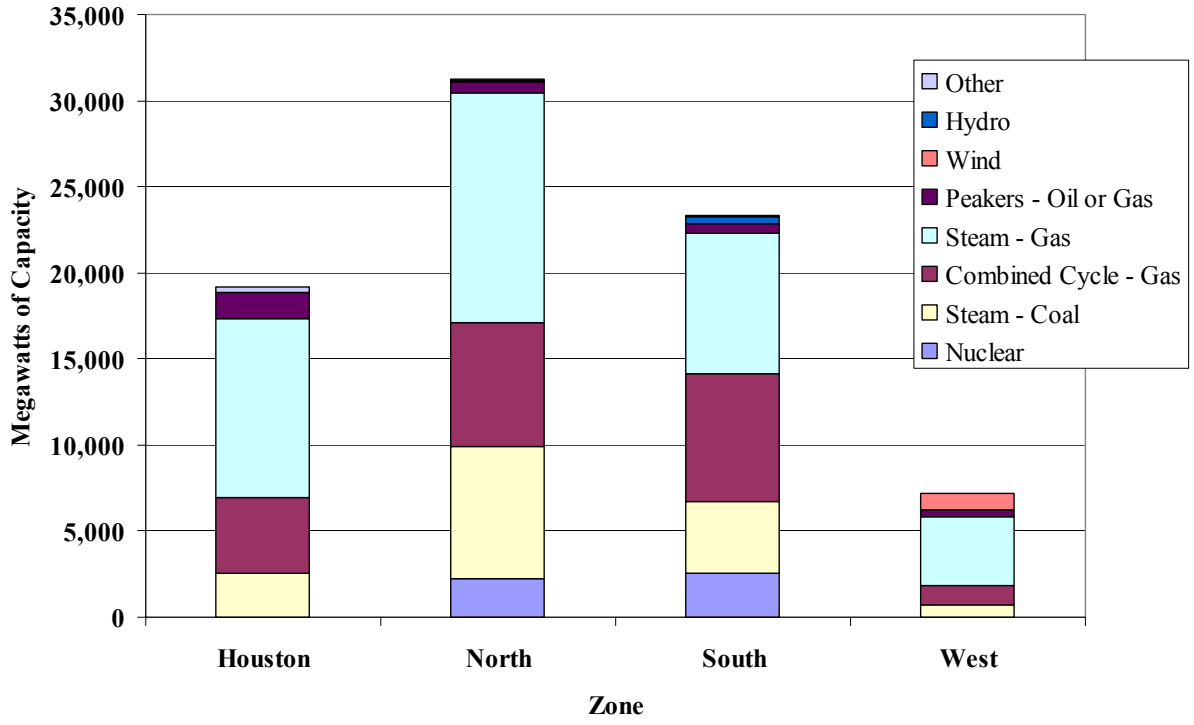
This figure also shows that the 60 GW peak load in 2003 was roughly 25 percent larger than the load at the 95th percentile of hourly load (approximately 48 GW). This is typical of the load patterns in an electricity market. Given that an additional 3,500 MW to 5,000 MW are needed to supply operating reserves and regulation, this implies that in long-run equilibrium with no surplus capacity, almost one-third of the generating resources are needed to supply energy in less

than 5 percent of the hours while maintaining required regulation and operating reserves.[20] This serves to emphasize the importance of efficient pricing during peak demand conditions to send accurate economic signals for the investment in and retention of these resources.

B. Generation Capacity in ERCOT

In this section we evaluate the generation mix in ERCOT. With the exception of the wind resources in the West Zone and the nuclear resources in the North and South Zones, the mix of generating capacity is relatively uniform in ERCOT. Figure 44 shows the generating capacity by type in each of the ERCOT zones.

Figure 44: Installed Capacity by Technology for each Zone



This figure shows that there is some nuclear capacity in both the North and South Zones, while lignite coal is also a major contributor in the North Zone. However, the primary fuel in all four zones is natural gas -- accounting for 73 percent of generation capacity in ERCOT as a whole,

[20] The range in the operating reserve and regulation requirements is based on the variable nature of the non-spinning reserves requirements.

and 85 percent in the Houston Zone. Much of this natural gas-fired capacity represents relatively new combined-cycle units than have been installed throughout ERCOT over the past few years.

ERCOT's reliance on natural gas resources makes it vulnerable to natural gas price spikes because coal and nuclear plants are primarily base load units. There is approximately 20,000 MW of coal and nuclear generation in ERCOT. Because there were only 23 hours during 2003 when ERCOT load was less than 20,000 MW, natural gas resources will be dispatched and set the balancing energy spot price in most hours. Hence, although coal-fired and nuclear units produce more than half of the energy in ERCOT, they play a much less significant role in setting spot electricity prices due to their relatively low marginal production costs.

The distribution of capacity among the ERCOT zones is similar to the distribution of demand. This is consistent with the legacy of investment under the regulated vertically integrated utilities when load and resources were integrated in independent geographic areas. The North Zone accounts for 39 percent of capacity, the South Zone 29 percent, the Houston Zone 24 percent, and the West Zone 9 percent. The ratio of generating resources to load is slightly higher in the South and lower in Houston than the ERCOT average. This helps explain the patterns of exports from the South to Houston, as discussed below.

1. ERCOT Resource Margins

In this subsection, we estimate the resource margin in ERCOT based on the actual peak demand and installed capacity over the past two years. A resource margin indicates the amount of generating resources (including imports) that are available in excess of the peak load as a percentage of the peak load. Table 4 provides a detailed breakdown of generation capacity by technology type and resource margins in ERCOT.

This table shows that ERCOT had substantial excess capacity in 2002 and 2003. Excluding mothballed capacity and import capability, resource margins for ERCOT as a whole have remained above 20 percent the last two years. When import capability from external ties and switchable resources are included, the resource margin rises to 31 percent in 2003. When including total potential response from loads acting as resources, the resource margin rises to 33 percent in 2003. This is particularly notable given that the 2003 peak demand level was approximately 2 GW higher than the planning forecast level (i.e., the resource margin based on

the planning forecast would have been significantly higher). At the zonal level, resource margins remain above 20 percent for all four zones, and are as high as 37 percent in the South Zone and 45 percent in the West Zone.

Table 4: Generation Capacity and Resource Margins in ERCOT

Category		Formula	2002	2003
Installed Capability by Type (MW)				
Nuclear			4,737	4,737
Steam - Coal			16,021	15,133
Combined Cycle - Gas			13,420	17,111
Steam - Gas			35,943	35,943
Peakers - Oil or Gas			3,026	3,026
Wind(10% included here)			74	94
Hydro			552	552
Other			413	413
Total Capacity	(1)		74,186	77,009
Out-of-Service Capacity (MW)				
Mothballed Capacity	(2)		5,527	2,420
In-Service Capacity	(3)	= (1) - (2)	68,659	74,589
Imported Capacity (MW)				
Switchable Capacity	(4)		1,726	3,068
DC Tie Import Capacity	(5)		856	856
In-Service Capacity Incl. Imports	(6)	= (3) + (4) + (5)	71,241	78,513
LaaRs - Loads Acting as Resource	(7)		570	1,200
In-Service Capacity, Imports, LaaRs	(8)	= (6) + (7)	71,811	79,713
Actual Peak Demand (MW)	(9)		56,248	59,996
Ratio of Resources to Actual Peak Demand:				
No Imports, Switchable, LaaRs	(10)	= (3) / (9) - 1	**22%**	**24%**
Plus Switchable*	(11)	= (4) / (9) + (10)	**25%**	**29%**
Plus DC-Tie Imports	(12)	= (6) / (9) - 1	**27%**	**31%**
Plus LaaRs**	(13)	= (8) / (9) - 1	**28%**	**33%**

* Most comparable to ERCOT methodology for calculating resource margin.

** This resource margin is over-estimated to the extent that the peak demand was reduced by the deployment of LaaRs (since the true peak would have been higher and LaaRs are already counted as resources).

Although these resource margins are sizable, it is important to consider that electricity demand in Texas has been growing at a rapid pace. From 1994 to 2003 the coincident peak grew at an annual rate of 3.6 percent,[21] despite a significant decline in 2001 due to the economic recession. At this rate, it will take little more than three years to reduce the ERCOT resource margin to 15 percent with no new generation. It is also important to consider that a significant number of generating units in Texas will soon be reaching or are already exceeding their expected operating lives. Over 8,300 MW of generation capacity is at least 40 years old, and another 18,600 MW of generation is between 30 and 40 years old.[22] Hence it is important to ensure that the ERCOT markets are designed to send efficient economic signals so that new investment occurs to maintain adequate resources as load grows and older resources retire.

2. Generation Outages and Deratings

The prior subsection shows substantial resource margins, indicating that the adequacy of resources is not a concern in ERCOT in the near-term. However, resource adequacy must be evaluated in light of the resources that are actually available on a daily basis to satisfy the energy and operating reserve requirements in ERCOT. A substantial portion of the installed capability is frequently unavailable due to generator deratings. A derating is the difference between a generating resource's installed capability and its maximum capability (or "rating") in a given hour. Generators can be fully derated (rating equals 0) due to a forced or planned outage. However, it is very common for generators to be partially derated (e.g., by 5 to 10 percent) because the resource cannot achieve its installed capability level due to technical factors or environmental factors (e.g., ambient temperature conditions).

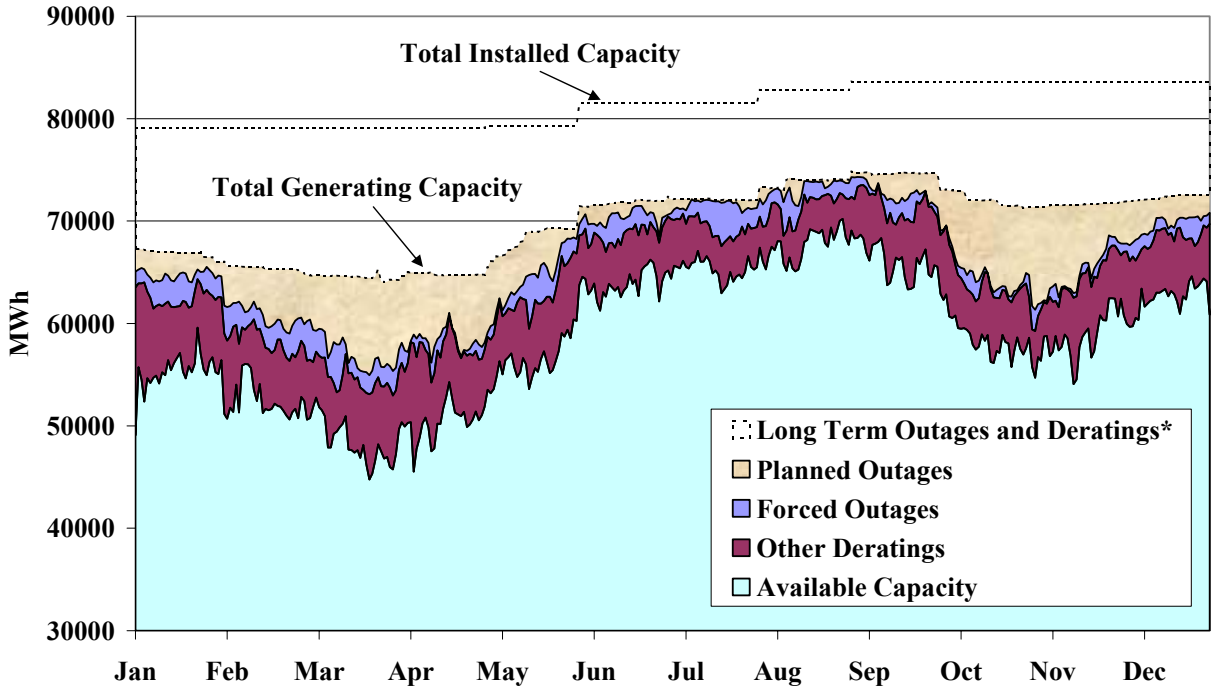
In this subsection, we evaluate long-term and short-term deratings to inform our evaluation of ERCOT capacity levels. Figure 45 below shows a breakdown of total installed capability for ERCOT on a daily basis during 2003. This analysis includes all in-service and switchable capacity. The capacity in this analysis is separated into five categories: (a) long-term outages and deratings, (b) short-term planned outages, (b) short-term forced outages, (c) other short-term deratings, and (d) available and in-service capability.

[21] ERCOT Transmission Study, 2003, p. 56.

[22] ERCOT Transmission Study, 2003, p. 69.

The long-term deratings category includes any outages and deratings lasting for 60 days or longer while the remaining outages and deratings are included in the short-term categories. We generally separate the long-term outages because it provides an indication of the generating capacity that is generally not available to the market, which typically exceeds 10 GW. Some of this capacity may be out-of-service for extended periods due to maintenance requirements or may be out-of-service during the spring and fall months for economic reasons. However, a large share of these deratings reflect output ranges on generating units that are not capable of producing up to the full installed capability level.

Figure 45: Short and Long-Term Deratings of Installed Capability 2003



**Includes all outages and deratings lasting greater than 60 days and all mothballed units*

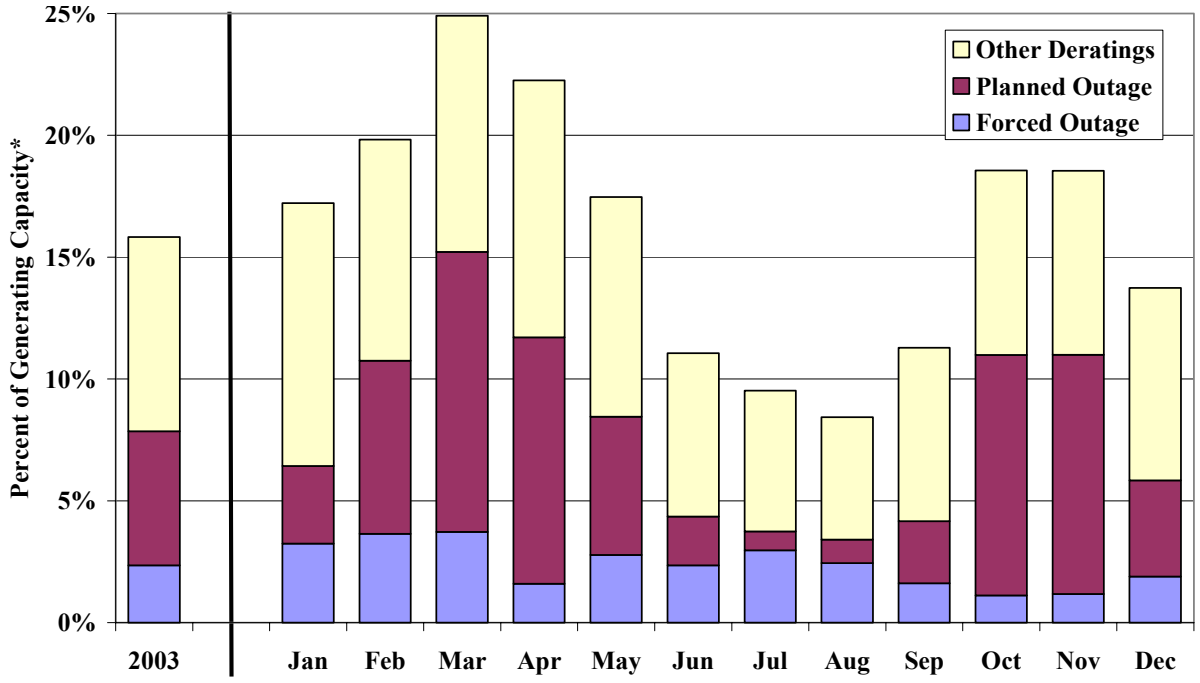
Figure 45 shows that installed capacity, including mothballed and switchable capacity, rose from just above 79 GW at the beginning of 2003 to more than 82 GW at the end of 2003. This increase is due to new capacity that became operable in 2003. The figure shows that the long-term outages and deratings fluctuated between 9 GW during the summer months and 14 GW during the spring and fall. This fluctuation is likely due to the fact that some of the more costly capacity may not be economic to keep in operation during the spring and fall. The long-term

outages and deratings also include 2,420 MW of mothballed capacity. These classes of capacity can be made available if market conditions become tighter as load rises.

As expected, planned outages are relatively large in the spring and fall, decreasing to close to zero during the summer. Available in-service capacity fluctuated between 44 GW in March and 69 GW in August. The peak hour for the year required 60 GW to satisfy ERCOT's energy requirements and an additional 3.5 to 5 GW for operating reserves and regulation-up requirements, resulting in surplus capacity of less than 5 GW. This surplus is much smaller than the resource margin statistics would imply.

The next analysis focuses specifically on the short-term outages and deratings. To more clearly show the outages and deratings lasting less than 60 days, Figure 46 shows the average magnitude of the outages and deratings lasting less than 60 days for the year and for each month during 2003.

Figure 46: Monthly Average Outages and Deratings*
2003



**Excludes all outages and deratings lasting greater than 60 days and all mothballed units*

Figure 46 shows that short-term deratings and outages were as large as 25 percent of installed capacity in the spring, dropping below 10 percent for most of the summer. Most of this fluctuation was due to anticipated planned outages, which ranged from approximately 10 to 11 percent of installed capacity during March, April, October, and November. Short-term forced outages occurred more randomly, as expected, ranging between 1 percent and 4 percent of total capacity on a monthly average basis during 2003. These rates are relatively low in comparison to other operating markets, which can be attributed to a number of factors mentioned below.

First, these outages include only full outages (i.e., where the resource's rating equals zero). In contrast, an equivalent forced outage rate is frequently reported for other markets, which includes both full and partial outages. Hence, the forced outage rate shown in Figure 46 can be expected to be lower than equivalent forced outage rates of other markets. Second, we were not confident that the forced outage logs received from ERCOT included all forced outages that actually occurred. Lastly, the largest category of short-term deratings was the "other deratings", which occur for a variety of reasons.

The other deratings would include any short-term forced or planned outage that was not reported or correctly logged by ERCOT. This category also includes natural deratings due to ambient conditions and other factors described above. Because these natural deratings can fluctuate day to day or seasonally, some of the deratings are included in the "long-term outages and deratings" category while the others are included in this category. The other deratings were approximately 5 percent on average during the summer in 2003 and as high as 10 percent in other months.

3. Daily Generator Commitments

One of the important characteristics of any electricity market is the extent to which it results in the efficient commitment of generating resources. Under-commitment can cause apparent shortages in real-time and inefficiently-high energy prices while over-commitment can result in excessive start-up costs, uplift charges, and inefficiently-low energy prices.

This subsection evaluates the commitment patterns in ERCOT by examining the levels of excess capacity. Excess capacity is defined as the total online capacity plus quick-start units minus the demand for energy, operating reserves, and up regulation. If the goal were to have no excess capacity, ERCOT would have to dispatch quick-start resources each day to meet its energy

demand.[23] Normally, however, because it is uneconomic to dispatch quick-start units for energy on most days, additional slow-starting resources with lower production costs are committed instead.

To evaluate the commitment of resources in ERCOT, Figure 47 plots the excess capacity in ERCOT during 2003. The figure shows the excess capacity in only the peak hour of each day because the commitments of generating resources are intended to cover the forecasted peak for the following day. Hence, one would expect larger quantities of excess capacity in other hours.

Figure 47: Excess Capacity
During Daily Peaks on Weekdays -- 2003

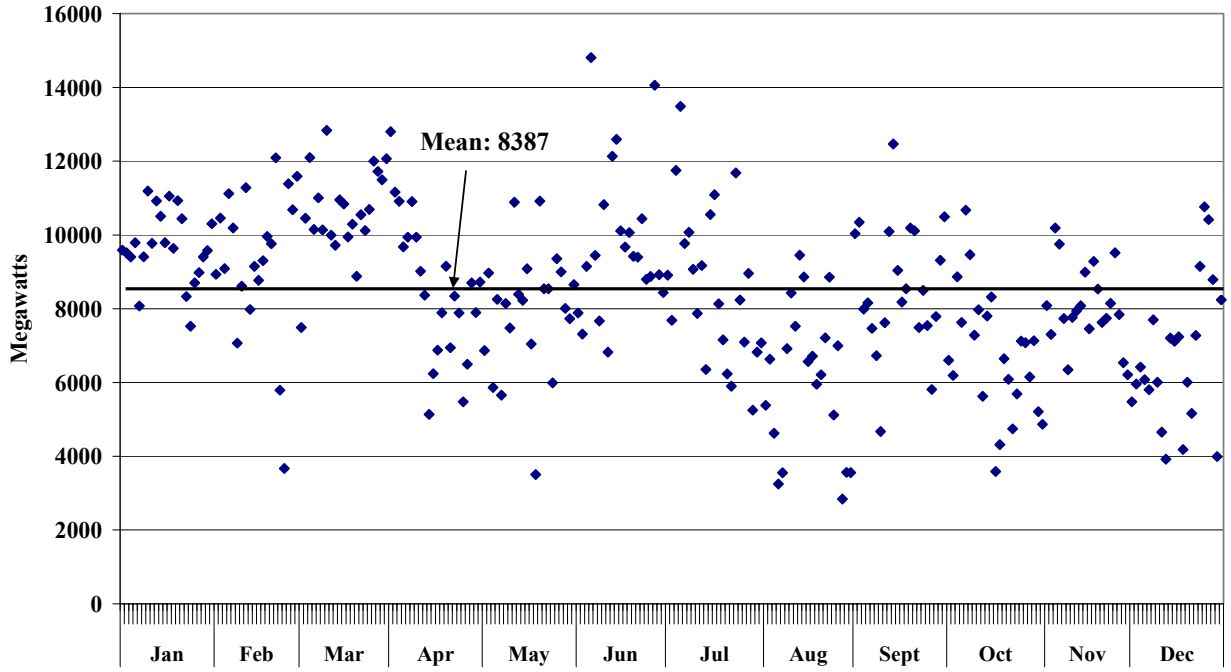


Figure 47 shows that the excess capacity in ERCOT was significant in 2003. The levels rarely fell below 4,000 MW on any day (and this occurred on only 9 days). During the peak load day in 2003 (on August 7), there were 3,550 MW available. However, the excess capacity averaged almost 8,400 MW, which is more than 20 percent of the average load in ERCOT. This is a sizable quantity of excess capacity.

[23] This is because there is about 3,000 MW of quick-start capacity of which only a portion can be used as reserves. Therefore, to have zero excess capacity in any day, all 3,000 MW of quick-start capacity would have to be committed and a portion of it would have to be used to supply energy.

Some of this excess capacity may be explained by the quick-start resources in ERCOT, of which there are approximately 3,000 MW available on average. However, even if all of these resources were excluded from the excess capacity calculation, the average excess capacity would still exceed 5,000 MW. The fact that the quantity of capacity committed exceeds the energy and ancillary services requirements by such a wide margin indicates that the current ERCOT market design tends to result in an over-commitment of resources. While this assists in ensuring reliability, this level of committed capacity is not efficient because these sizable excess resource commitments are costly.

The tendency to over-commit capacity can be attributed in large part to the lack of a centralized day-ahead commitment process in ERCOT. Without a centralized commitment mechanism, each participant makes independent generator commitment decisions that, taken together, are not likely to be optimal. Hence, the introduction of day-ahead energy and operating reserves markets under the Texas Nodal market design currently being developed promises substantial efficiency improvements in the commitment of generating resources.

C. Demand Response Capability

Demand response is a term that broadly refers to actions that can be taken by end users of electricity to reduce load in response to instructions from ERCOT or in response to certain market conditions. The ERCOT market allows participants with demand-response capability to provide the energy, reserves, and regulation in a manner similar to a generating resource. The ERCOT Protocols allow for loads to participate in the ERCOT administered markets as either Loads acting as Resources ("LaaRs") or Balancing Up Loads ("BULs").

ERCOT allows LaaRs that are qualified to offer responsive reserves, non-spinning reserves, or regulation into the day-ahead ancillary services markets. Those that are qualified can also offer blocks of energy in the balancing energy market. LaaRs providing up balancing energy must have telemetry and must be capable of responding to ERCOT energy dispatch instructions in a manner comparable to generation resources. Those providing responsive reserves must have high set under-frequency relay ("UFR") equipment. A load with UFR equipment is automatically tripped when the frequency falls below 59.7 Hz. LaaRs that are capable of controllably reducing or increasing consumption under dispatch control (similar to AGC) may

also provide regulation service. However, there are currently no such resources registered in ERCOT.

BULs are loads that are qualified to offer demand response capability in the balancing energy market. These loads must have an Interval Data Recorder to qualify and do not require telemetry. BULs may provide energy in the balancing energy market. Unlike LaaRs, however, they are not qualified to provide reserves or regulation service.

Table 5 provides details on participation in the ERCOT demand response programs in 2003. This table shows that there were 35 participants qualified as LaaRs with a total capability of 1,200 MW. No BULs were registered with ERCOT in 2003.

Table 5: LaaRs Participation in ERCOT Responsive Reserve Market

# of Registered Participants	35
Total Capability (MW)	1200
Average Responsive Reserves Bid (MW)	450
Average Responsive Reserves Struck (MW)	450
Average Hourly Self Provided (MW)	350
Total Average Hourly Participation (MW)	800

This table shows that LaaRs provided an average of 800 MW of responsive reserve to the market in 2003, divided almost equally between self-supplied reserves and offers cleared in the responsive reserves market. As of 2004, LaaRs are permitted to supply up to 1,150 MW of the responsive reserves requirement. The fact that all responsive reserves offered are procured, indicates that the LaaRs are offered at prices that are relatively low, causing them to virtually always clear in the responsive reserves market. The total quantity of responsive reserves supplied by LaaRs represented 35 percent of the total 2,300 MW requirement for responsive reserves in 2003, and will likely represent 50 percent of this requirement in 2004, which sets ERCOT apart from other operating electricity markets. LaaRs serving as responsive reserves were deployed twice in 2003 when they were curtailed automatically through the UFRs.

Although LaaRs are active participants in the responsive reserves market, they did not provide offers in the balancing energy, non-spinning reserves, or regulation services markets in 2003. This is not surprising because the value of curtailed load (or the cost of dispatching distributed

generation) tends to be relatively high and providing responsive reserves can earn substantial revenue with very little probability of being deployed. In contrast, providing non-spinning reserves introduces a much higher probability of being curtailed. Participation in the regulation services market requires technical abilities that LaaRs cannot meet at this point. Finally, prices in the balancing energy market have not been high enough to attract load participation in that market. Hence, most LaaRs will have a strong preference for providing responsive reserves over regulation services, spinning reserves, or balancing energy.

IV. TRANSMISSION AND CONGESTION

One of the most important functions of any electricity market is to manage the flows of power over the transmission network by limiting additional power flows over transmission facilities when they reach their operating limits. In ERCOT, constraints on the transmission network are managed in two ways. First, ERCOT is made up of zones with the constraints between the zones managed through the balancing energy market. The balancing energy market increases energy production in one zone and reduces it in another zone to manage the flows between the two zones when the interface constraint is binding i.e., when there is interzonal congestion. Second, constraints within each zone (i.e., local congestion) are managed through the redispatch of individual generating resources. In this section of the report, we evaluate the ERCOT transmission system usage and analyze the costs and frequency of transmission congestion.

A. Electricity Flows between Zones

In 2003, there were four commercial pricing zones in ERCOT: (a) the North Zone, (b) the West Zone, (c) the South Zone, and (d) the Houston Zone. The balancing energy market uses the SPD software that dispatches balancing energy in each zone in order to serve load and manage congestion between zones. The SPD model embodies the market rules and requirements documented in the ERCOT protocols.

To manage interzonal congestion, SPD uses a simplified network model with four zone-based locations and three transmission interfaces. These three transmission interfaces, referred to as Commercially Significant Constraints ("CSCs"), are simplified representations of groups of transmission elements. ERCOT operators use planning studies and real-time information to set limits for each CSC that are intended to approximate the total transfer capability of the CSC. In this subsection of the report, we describe the SPD model's simplified representations of flows between zones and analyze actual flows in 2003.

The SPD uses zonal approximations to represent complex interactions between generators, loads, and transmission elements. Because the model flows are based on zonal approximations, the estimated flows can depart significantly from real-time physical flows. Estimated flows that diverge significantly from actual flows can result in inaccurate congestion modeling leading to

inefficient energy prices and other market costs. This subsection analyzes the impact of SPD transmission flows and constraints on market outcomes. A future report will examine in more detail ERCOT's market operations, including the extent and causes of divergence between SPD-modeled flows and physical flows. Figure 48 shows the average SPD-modeled flows over CSCs between zones during 2003.

Figure 48: Average SPD-Modeled Flows on Commercially Significant Constraints 2003

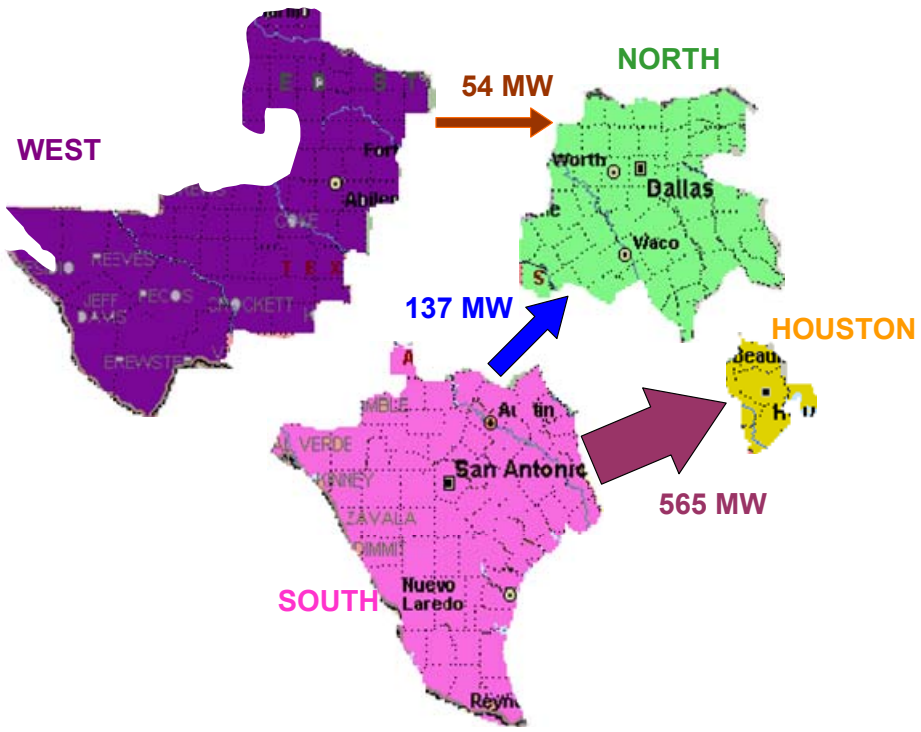


Figure 48 shows the four ERCOT geographic zones as well as the three CSCs that interconnect the zones: (a) the West to North interface, (b) the South to North interface, and (c) the South to Houston interface. The only zonal interface which exhibits substantial SPD-modeled energy flows is the South Zone to Houston Zone interface, with an average of 565 MW flowing across it into Houston. The South Zone exports an average of 137 MW to the North Zone. The West Zone exports 54 MW to the North Zone across the West to North interface. These results indicate that in general the four ERCOT zones function relatively independent of each other and do not rely heavily on imports from adjacent zones. This is consistent with the historical patterns

of supply in each area that developed in an era when most demand was served by the utilities' own generation.

As discussed above, the simplified modeling assumptions specified in the ERCOT protocols for the current zonal market causes the interzonal power flows calculated by SPD to frequently diverge significantly from the actual flows. The most important simplifying assumption is that all generators in a zone have the same effect on the flows over the CSC, or the same generation shift factor ("GSF")[24] in relation to the CSC. In reality, the generators within each zone can have widely varying effects on the flows over a CSC. In order to illustrate this, we calculated flows that would occur over the CSC using actual generation and actual generation shift factors and compared this to transmission flows calculated using actual generation and zonal average shift factors. Table 6 shows this analysis.

Table 6: Average Calculated Flows on Commercially Significant Constraints Zonal-Average vs. Unit-Specific GSFs – 2003

CSC	Flows Modeled by SPD (1)	Flows Calculated Using Actual Generation (2)	*Difference = (2) - (1)*	Flows Calculated Using Actual Generation and Unit-specific GSFs (3)	*Difference = (3) - (2)*
West-North	54	-10	*-64*	-157	*-147*
South-North	137	132	*-5*	285	*153*
South-Houston	565	518	*-47*	730	*212*

The first column in Table 6 shows the average flows over each CSC calculated by SPD. The second column shows the average flows over each CSC we calculated using zonal-average GSFs and actual real-time generation in each zone instead of the scheduled energy and balancing energy deployments used as an input in SPD. Although these flows are both calculated using the same zonal-average GSFs, they can differ when the actual generation varies from the SPD generation. This difference is shown in the third column (in italics). These differences indicate that the actual generation levels result in slightly lower calculated flows on each CSC. The fourth column in Table 6 reports the average flows over each CSC calculated using unit-specific

[24] A GSF indicates the portion of the incremental output of a unit that will flow over a particular transmission facility. For example, a GSF of 0.5 would indicate that half of any incremental increase in output from a generator would flow over the interface. Likewise, a GSF of -0.5 would indicate that an incremental increase of 1 MW would reduce the flow over the interface by 0.5 MW.

GSFs and actual real-time generation. Since the actual generation data used to calculate the flows in this column are identical to those used in column (2), the difference in flows between the two columns can be attributed to using zonal GSFs versus resource-specific GSFs. These differences in flows are shown in the fifth column (in italics). The differences in the last column measures the inaccuracy caused by treating each unit within a particular zone as having identical impacts on the CSCs.

These results show that the heterogeneous effects of generators in a zone on the CSC flows can cause the actual flows to differ substantially from the SPD-calculated flows. Table 6 shows that the unit-specific GSFs reduced the calculated flows on the West-North interface by 147 MW and increased the calculated flows on the South-North and South-Houston interfaces by 153 MW and 212 MW, respectively. These differences are sizable and are much larger than the differences that can be attributed to variations in actual generation.

We note that the GSF simplification embedded in the SPD model is important for loads as well. Loads tend to be concentrated within a zone, but the SPD model assumes a generation-weighted average shift factor for all loads in the zone. We did not have the data to evaluate the size of the CSC flow effect that this assumption creates, but this will be studied in more detail in the market operations report we will be issuing in the fall of 2004.

In order to effectively manage interzonal congestion, it is important for SPD to accurately model the major constrained transmission interfaces between zones. In 2003, the three CSCs modeled by SPD did not include all significant interfaces between zones, while substantial quantities of power were transported on transmission facilities not modeled by SPD. For example, the Houston Zone imported an average of 1,796 MW in 2003 although SPD only modeled an average import of 565 MW. Table 7 summarizes the actual net imports into each zone compared to SPD modeled flows.

Table 7: Actual Net Imports vs. SPD-Calculated Flows on CSCs 2003

Zone	Actual Net Imports	SPD Flows on CSCs
Houston	1796	565
North	507	191
South	-1213	-702
West	-76	-54

Table 7 illustrates how different the SPD-calculated flows are on average than the actual flows into each zone. These differences can be attributed to two factors. First, the use of zonal average GSFs by SPD can cause the SPD-calculated flows on a particular CSC to be substantially different from the actual flows. However, based on the previous analysis in Table 6, this cannot explain all of the difference between actual net interchange and interchange modeled on CSCs.

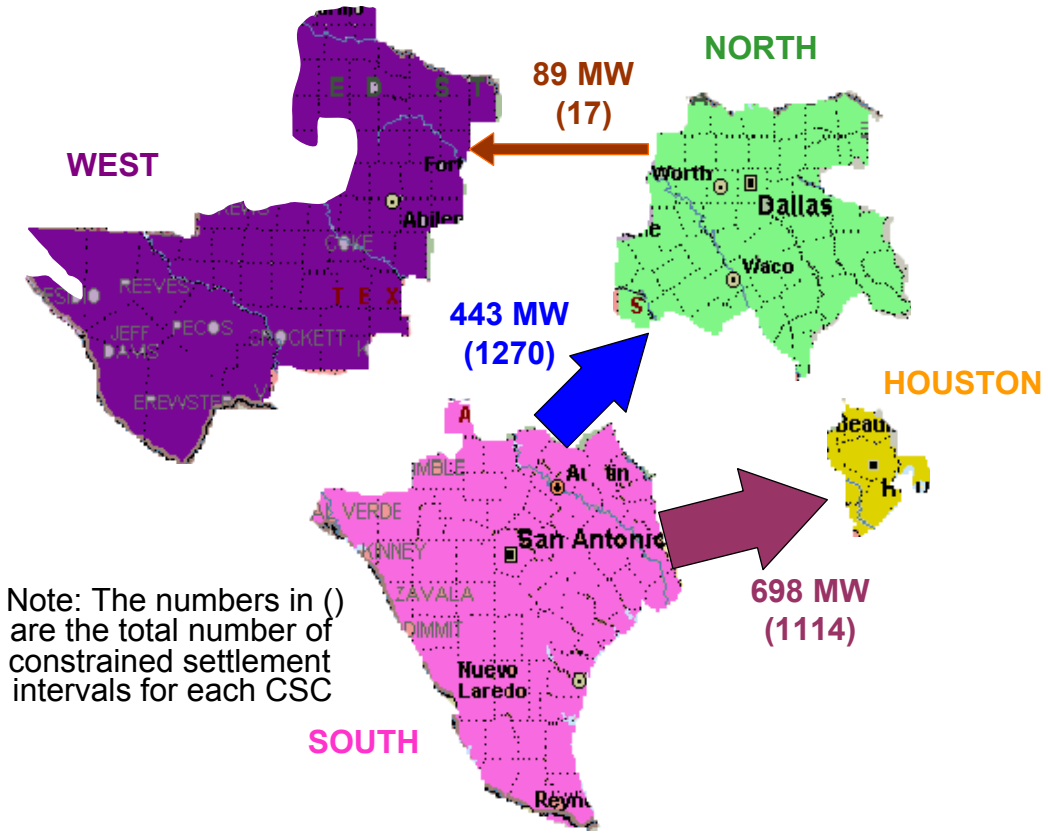
Second, significant quantities of power may flow over other transmission facilities that are not defined as part of the CSC. This will tend to cause the actual imports to exceed the SPD-calculated flows over the CSCs. Most notably, the Houston Zone imports large quantities from the North Zone that causes frequent congestion that is not managed by zonal balancing deployments. When these additional flows do not cause transmission constraints to bind, they raise no significant market issues. However, if transmission constraints between zones that are not defined as part of a CSC do become binding, ERCOT's means for managing the constraints can result in inefficiencies. In this case, it may be beneficial to redefine the CSC or define a new CSC that will include these interzonal flows. For example, ERCOT introduced the North to Houston CSC in 2004 to allow it to better manage interzonal congestion.

B. Interzonal Congestion

The prior subsection showed the average interzonal flows calculated by SPD compared to actual flows in all hours. This subsection focuses on those intervals when the interzonal constraints were binding. Although this is a small subset of intervals, it is in these constrained intervals that the performance of the market is most critical.

Figure 49 shows the average SPD-calculated flows between the four ERCOT zones during constrained periods for the three CSCs. The arrows show the average magnitude and direction of the SPD-calculated flows during constrained intervals. The frequency with which these constraints arise is shown in parentheses.

Figure 49: Average Modeled Flows in Transmission Constrained Intervals 2003



As Figure 49 shows, the SPD-calculated flows averaged 698 MW on the South to Houston interface during 1,114 constrained intervals in 2003. The average SPD-calculated flows on the South to North interface was 443 MW during the 1,270 intervals the interface was constrained. The South to Houston and South to North interfaces exhibited higher SPD-calculated flows in these constrained intervals than the average flows in the other intervals. On the West to North interface, the SPD-calculated flows averaged -89 MW from the West to the North Zone during the 17 constrained intervals. Since the transmission interface is defined as one that limits the flows from West to North, at least some of the 17 intervals had negative flow limits in SPD, meaning that SPD-calculated flows were constrained to be in the direction of the West Zone. In

these intervals, SPD would have allowed more power to flow into the West Zone, but not less. This particular result is examined in more detail below.

1. Congestion Rights

Interzonal congestion can be significant from an economic perspective, compelling the dispatch of higher-cost resources because power produced by lower-cost resources cannot be delivered over the constrained interfaces. When this occurs, participants must compete to use the available transfer capability between zones. In order to allocate this capability efficiently, ERCOT establishes clearing prices for energy in each zone that will vary in the presence of congestion and charges the transactions between the zones the difference in these prices. The zonal price difference – the congestion price charged to interzonal transactions – represents the marginal cost to the system of the constraint (i.e., the savings that would be achieved by increasing the transfer capability between the zones by one megawatt).

Market participants in ERCOT can hedge congestion costs in the balancing energy market by acquiring Transmission Congestion Rights ("TCRs") or Pre-assigned Congestion Rights ("PCRs"). Both TCRs and PCRs entitle the holder to payments corresponding to the interzonal congestion price. Hence, a participant holding TCRs or PCRs for a transaction between two zones would pay the interzonal congestion price associated with the transaction and receive TCR and/or PCR payments that fully offset the congestion charges. TCRs are acquired by annual and monthly auctions (as explained in more detail below) while PCRs are allocated to certain participants based on historical patterns of transmission usage.

In order to analyze the congestion rights in ERCOT, we first review the TCRs and PCRs that were allocated for each CSC in 2003. Figure 50 shows the average number of TCRs and PCRs that were allocated for each of the CSCs in 2003, as well as the average SPD-modeled flows during the constrained intervals.

Figure 50: Transmission Rights vs. Real-Time SPD-Calculated Flows Constrained Intervals 2003

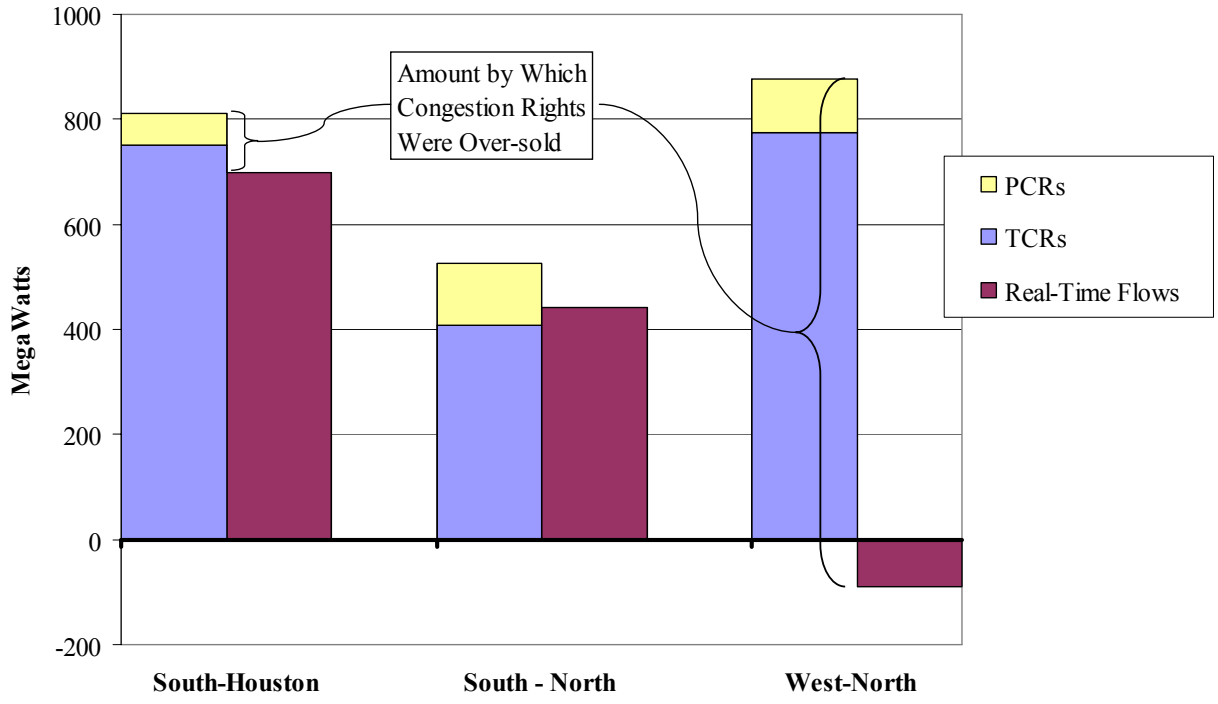


Figure 50 shows that total congestion rights (the sum of PCRs and TCRs) on the South to Houston interface exceeded the average real-time SPD-calculated flows during constrained hours by more than 100 MW. These results indicate that the congestion rights were oversold in relation to the SPD-calculated limits by about 100 MW on average. Likewise, the South to North CSC congestion rights were oversold by 83 MW.

The largest divergence between the SPD-calculated limits and the limits implied by the congestion rights was on the West to North interface where 877 MW of congestion rights were sold, but the average SPD-calculated flow during constrained intervals was actually negative 89 MW. Hence, the congestion rights that determine ERCOT's total obligation to make congestion payments exceeded the modeled flow over the CSC by nearly 1,000 MW. As explained in more detail below, the negative SPD-calculated flows from the West Zone to the North Zone are misleading. It is likely that real-time physical flows were actually positive in these intervals, diverging substantially from the SPD-calculated flows.

Ideally, the financial obligations to holders of congestion rights would be satisfied with congestion revenues collected from participants scheduling over the interface and through the sale of balancing energy that flows over the interface. When the SPD-calculated flows are consistent with the quantity of rights sold over the interface, the congestion revenues will be sufficient to satisfy the financial obligations to the holders of the congestion rights. Alternatively, when the quantity of congestion rights exceeds the SPD-calculated flow over an interface, the congestion revenues from the balancing energy market will not be sufficient to meet the financial obligations to congestion rights holders.

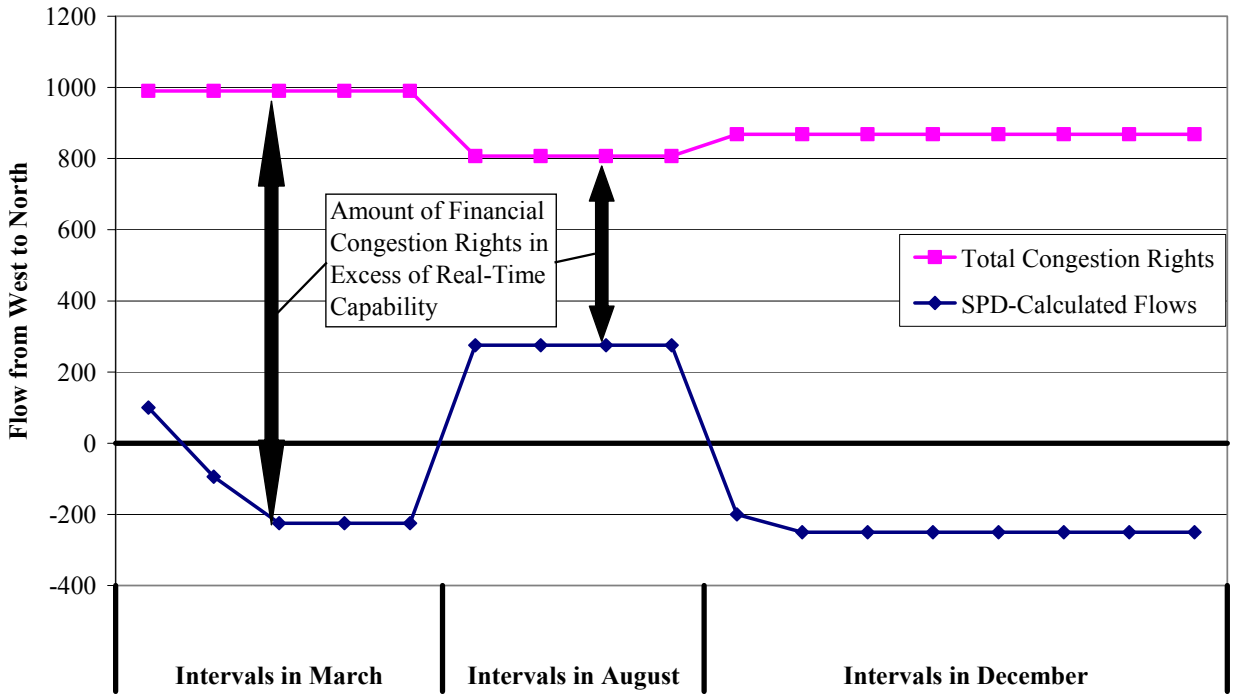
For instance, suppose the SPD-calculated flow limit is 300 MW for a particular CSC during a constrained interval. Also suppose that the holders of congestion rights own a total of 800 MW over the CSC. ERCOT will receive congestion rents from the balancing energy market that cover precisely 300 MW of the 800 MW worth of obligations. Thus, a revenue shortfall will result that is proportional to the shadow price of the constraint on the CSC in that interval (i.e., proportional to the congestion price between the zones). In this case, the financial obligations to the congestion rights holders cannot be satisfied with the congestion revenue, so the shortfall is charged proportionately to all loads in ERCOT as part of the Balancing Energy Neutrality Adjustment charges.

To better understand the nature and causes of the shortfall implied by the results of Figure 50, we compare the SPD-calculated flows and congestion rights quantities for each of the constrained intervals by CSC.

2. West to North Interface

The first CSC we analyze at the interval level is the West to North CSC for which Figure 50 shows counterintuitive results. Figure 51 shows the total quantity of congestion rights allocated by ERCOT for the West to North interface relative to the real-time SPD-calculated flows over the interface in all 17 intervals that were constrained during 2003. In the figure, Total Congestion Rights include both TCRs and PCRs.

Figure 51: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals West to North Interface – 2003



There were five intervals in March, four in August, and eight in December when the West to North interface was constrained. Only five of the constrained intervals actually exhibited SPD-modeled flows in West to North direction, while the other twelve exhibited modeled flows in the "counterintuitive" North to West direction. The SPD-calculated flow was closer to the forecast transmission limit (on which TCR sales are based) in August, when the difference between the two was 532 MW. The SPD-calculated flow differed from the forecast transmission limit the most in March, when the difference was 1,215 MW. A congestion revenue shortfall was generated in all 17 intervals.

Although congestion was extremely infrequent on the West to North interface during 2003, balancing energy market outcomes were highly anomalous during intervals when the interface was constrained. These anomalies occur when the modeled flows in SPD are substantially different from actual physical flows in real time, i.e., when the actual system conditions result in more flows over the West to North constraint than the simplified zonal model would predict. To prevent physical flows from exceeding the physical limits of the CSC, the ERCOT operators

manually reduce the limit on the West to North interface in SPD. This causes SPD to redispatch generation in the various zones to reduce flows over the interface. Hence, because the SPD-calculated flows can be substantially different than actual flows, the ERCOT operators manage congestion by lowering the SPD limit when a constraint is physically binding to prevent additional flow over the CSC.

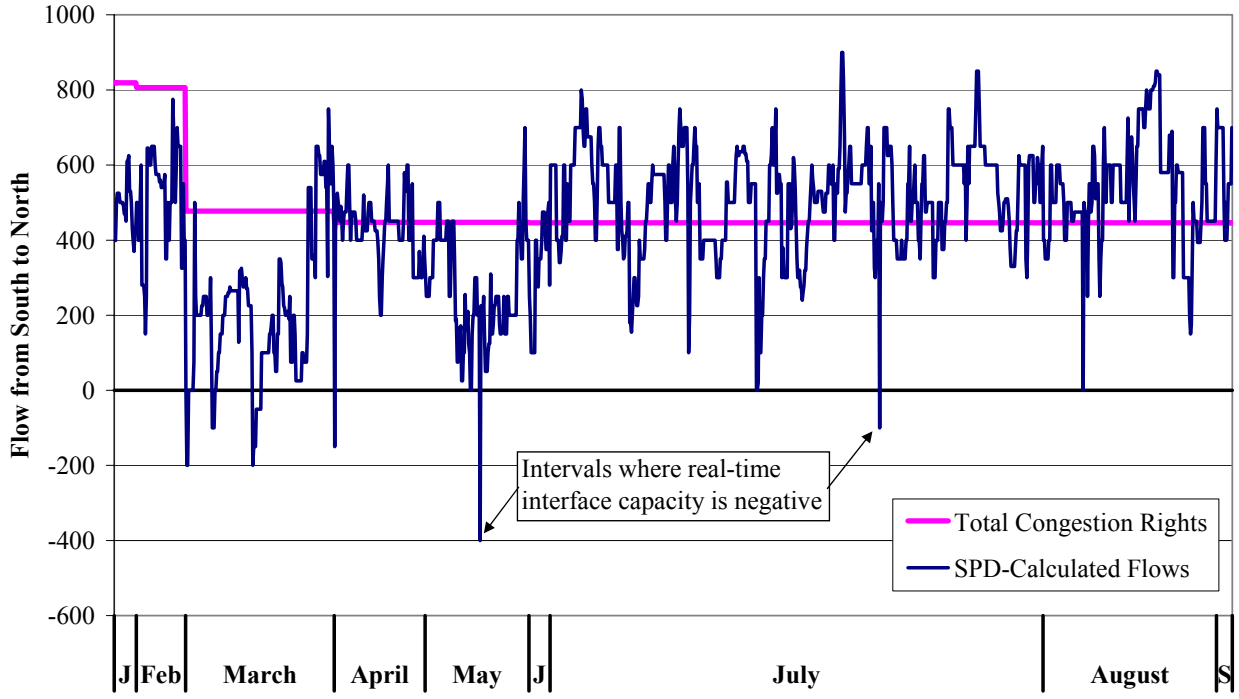
Even though the figure shows the SPD flows departed significantly from congestion rights quantities, it is likely that the actual physical flows were closer to the congestion rights quantities. In this case, the actual physical flows in each of the constrained intervals are from West Zone to the North Zone. However, the SPD-calculated flows are so different than the actual flows at times that the ERCOT operators must set a negative limit for the interface, which creates the apparent flow from the North to the West Zone.

3. South to North Interface

None of the other CSCs exhibit results as anomalous as the West to North CSC, although they do show substantial differences between congestion right quantities and SPD-calculated flows at the interval level. Figure 52 shows the total quantity of congestion rights allocated by ERCOT for the South to North interface relative to the real-time SPD-calculated flows over the interface when the constraint was binding during 2003. Because only congested intervals are shown, some months will have significantly more observations than other months. Indeed, the figure shows that congestion occurred with moderate frequency in January through June while July and August accounted for more than half of all constrained hours during 2003.

As explained in more detail below, the projected quantity of congestion rights changes from month to month as ERCOT reassesses the capability of each interface. ERCOT then adjusts the quantity of TCRs accordingly in the monthly auctions. Figure 52 shows these changes in the congestion rights relative to the SPD-calculated flows, which fluctuate considerably in the congested intervals.

Figure 52: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to North 2003



From January to March and in May, the SPD flows in the congested hours were almost always lower than the total quantity of congestion rights sold. During April and the summer months, the quantity of congestion rights more closely matched the average SPD-calculated flows over the CSC. The figure also shows specific instances where the SPD-calculated flows were constrained as low as negative 400 MW. Some of these instances of reduced SPD-calculated flows may be attributable to transmission outages. However, a significant share of the relatively low SPD-calculated flows corresponds to periods of unusually large differences between actual physical flows and the SPD-calculated flows, particularly in cases where the SPD-calculated flows were extremely low or negative.

4. South to Houston Interface

Figure 53 shows the total quantity of congestion rights allocated by ERCOT for the South to Houston interface relative to the SPD-calculated flows over the interface in congested hours during 2003.

Figure 53: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to Houston 2003

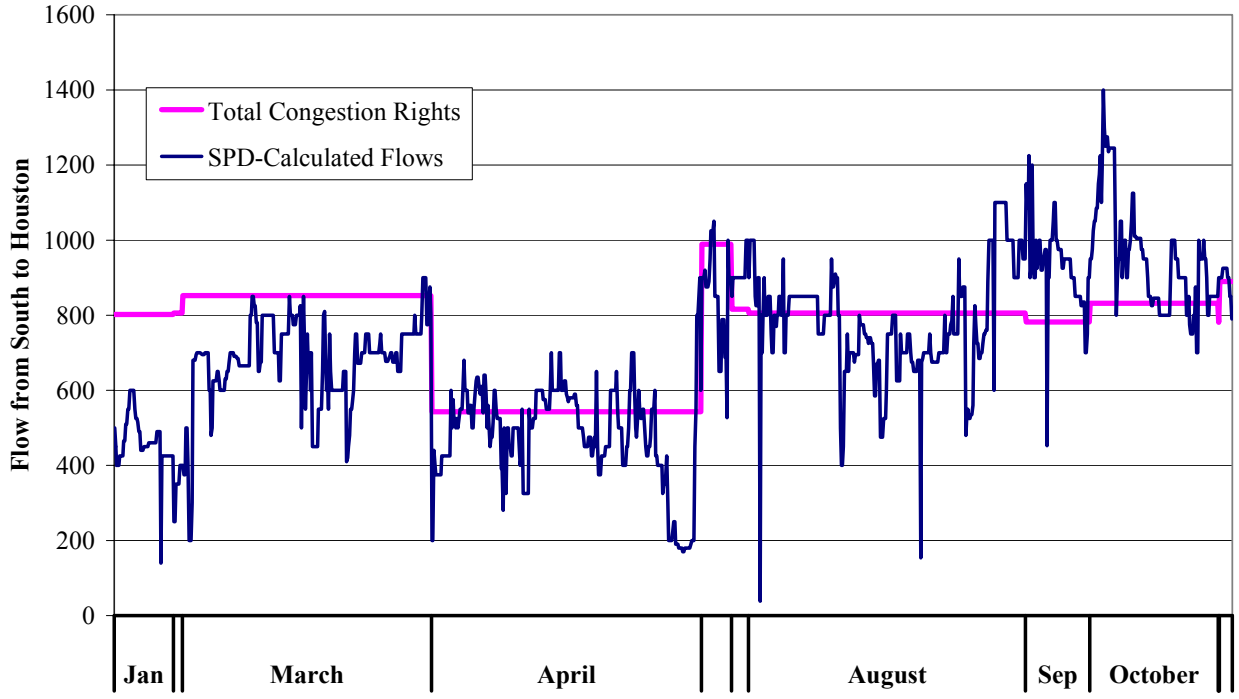


Figure 53 shows that the quantity of congestion rights on the South to Houston interface was approximately 800 MW at the beginning of 2003. The quantity of congestion rights was reduced to approximately 550 MW during the month of April, consistent with the reduction in SPD-calculated flows during this period. During the summer, congestion rights slightly exceeded SPD flows during constrained intervals. However, SPD flows during constrained intervals in the fall exceeded the allocations of congestion rights.

Transmission outages can have a significant effect on these results by reducing the flows that will be allowed by SPD. When the outage is recognized prior to when the monthly congestion rights are sold, ERCOT will reduce the quantity of rights that are made available to participants, which would prevent the outage from causing a significant shortfall associated with a large divergence between the congestion rights and the SPD flows. However, short-term outages that are not recognized in the monthly auctions can contribute to such divergences and result in revenue shortfalls. The next section describes ERCOT's process for selling congestion rights and reviews the results of these sales for 2003.

In conclusion, the SPD-calculated flows can vary substantially and frequently they are not close to the actual flows or limits for the CSC. Because transmission rights are generally sold based on the actual CSC transfer capability, this can result in substantial surplus congestion revenue or congestion revenue shortfall that results in uplift charges. Under the current market design, it is extremely difficult to develop procedures for selling transmission rights that fully subscribe the available transmission capability.

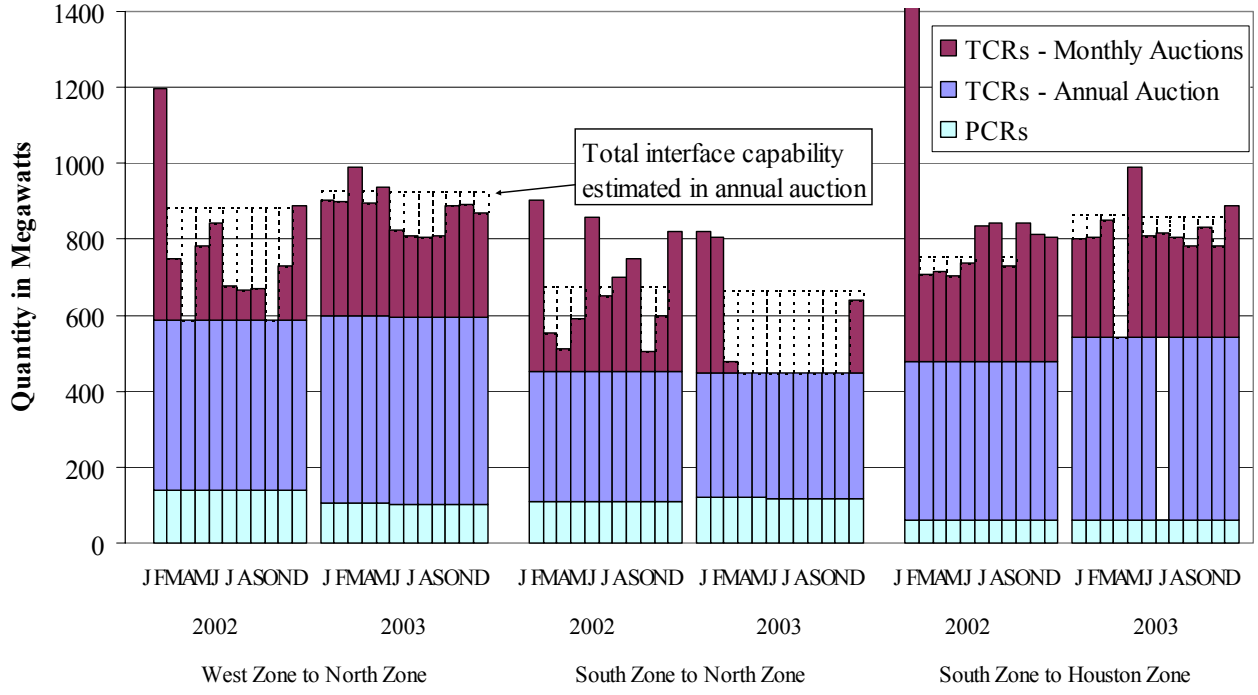
C. Congestion Rights Market

In this subsection, we review ERCOT's process to establish the quantity of congestion rights allocated or sold to participants. ERCOT performs transmission planning studies to determine the capability of each interface under peak summer conditions. This summer planning study is the basis for designating 60 percent of the congestion rights sold in the annual auction. These rights are auctioned in December for the coming year. The remaining 40 percent of the rights are designated based on monthly updates of the summer study. Since the monthly studies tend to more accurately reflect conditions that will prevail in the coming month, the monthly designations tend to more closely reflect actual transmission limits.

However, the summer and monthly studies used to designate the TCRs do not reflect transmission conditions that can arise in real-time. This happens for two main reasons. First, transmission and generator outages can occur unexpectedly, and significantly reduce the transfer capability of a CSC. Second, conditions may arise that cause the actual physical flow to be significantly different from the SPD modeled flow. As discussed above, ERCOT operators may need to respond by lowering the SPD-modeled flow limits in order to manage the actual physical flow. Accordingly, it is likely that the quantity of congestion rights will be larger than available transmission capability in SPD. This is one potential source of divergence for the West-North interface shown above.

To examine how these processes have together determined the total quantity of rights sold over each interface, Figure 54 shows the quantity of each category of congestion rights for each month since February 2002. The quantities of PCRs and annual TCRs are constant across months and determined before the beginning of each year, while monthly TCR quantities can be adjusted monthly.

Figure 54: Quantity of Congestion Rights Sold by Type
2002 v 2003



When the monthly planning studies indicate changes from the summer study, revisions are often made to the estimated transmission capability. Therefore, the auctioned congestion rights may increase or decrease relative to the amount estimated in the summer study. The shadow boxes in the figure represent the capability estimated in the summer study that is not ultimately sold in the monthly auction. When there is no shadow box in Figure 54, the total quantity of PCRs and TCRs sold in the annual and monthly auctions equaled or exceeded the summer estimate and therefore no excess capability is shown.

The South to North interface experienced the most fluctuation in the estimates of transmission capacity from the annual auction to the monthly auction. In fact, South to North TCRs were not even auctioned from April to November 2003 in the monthly auctions. The divergence between annual and monthly estimates of transmission capacity on the other interfaces was less extreme. The annual values and the monthly values on the West to North interface, in particular, were closely matched in 2002 and 2003.

Market participants who are active in congestion rights auctions are subject to substantial uncertainty. Outages and other contingencies occur randomly that can substantially change the

market value of a congestion right. Real-time congestion prices reflect the cost of interzonal congestion and are the basis for congestion payments to congestion rights holders. In a perfectly efficient system with perfect forecasting by participants, the average congestion price should equal the auction price. However, we would not expect full convergence in the real-world, given uncertainties and imperfect information. To evaluate the results of the ERCOT congestion rights market, in Figure 55 we compare the annual auction price for congestion rights, the average monthly auction price for congestion rights, and the average congestion price for each CSC.

Figure 55: TCR Auction Prices versus Balancing Market Congestion Prices 2002 v 2003

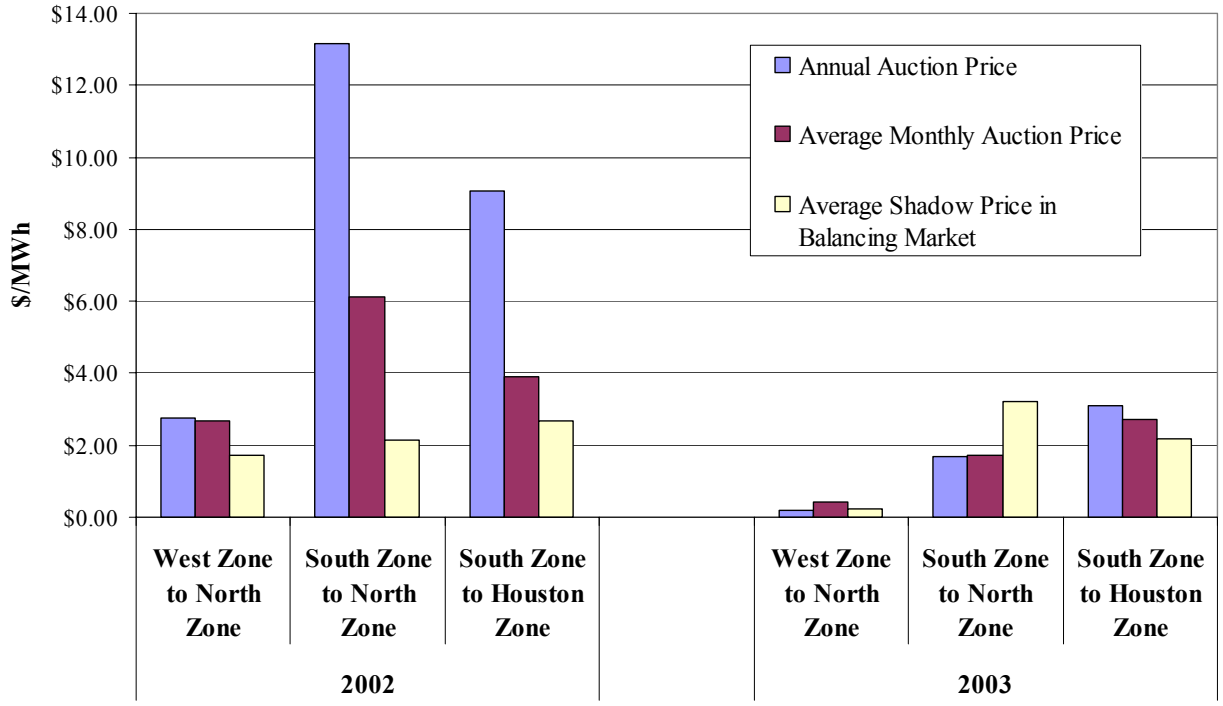
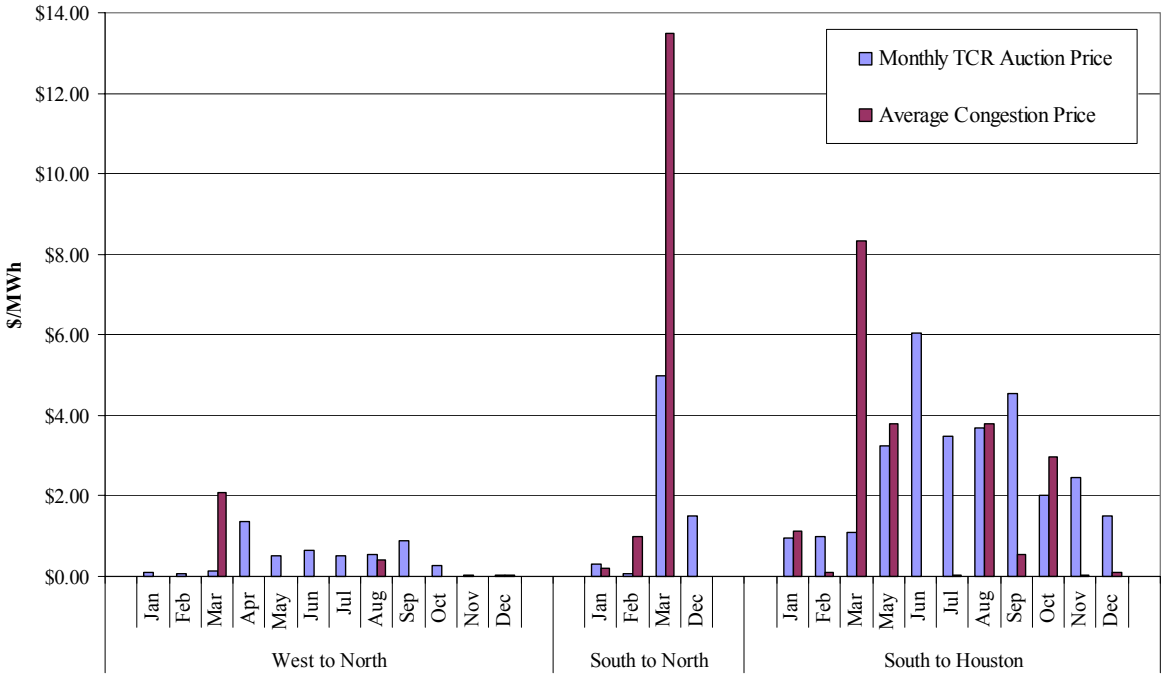


Figure 55 shows a substantial change from 2002 to 2003. In 2002, the annual auction for the TCRs resulted in prices that substantially over-valued the congestion rights on the South to North and South to Houston interfaces. Monthly TCR prices for these interfaces were roughly one-half of the prices from the annual auctions, but were still significantly higher than the ultimate congestion payments to the TCR holders. In the West to North interface, the annual and monthly TCR auction prices were close in magnitude and were both much closer to the true value of the congestion rights.

In 2003, the TCR prices for all of the interfaces decreased considerably, causing the prices to converge more closely with the actual value of the congestion rights. This indicates that participants have improved in their ability to forecast interzonal congestion and to value the TCRs. This improvement is likely facilitated by the simplified zonal representation of the ERCOT network embedded in the balancing energy market.

Figure 56 compares monthly TCR auction prices with monthly average real-time CSC shadow prices from SPD for 2003. To compare these more easily, the TCR auction prices are expressed in dollars per MWh. In months when the monthly auction did not occur (i.e., when the annual auction designated sufficient congestion rights for that month) no data is presented. This explains the missing months for the South-North interface and the South-Houston interface.[25]

Figure 56: Monthly TCR Auction Price and Average Congestion Value 2003



For the West to North interface, Figure 56 shows that the average real-time congestion price was positive in only three months. However, because the TCR auction prices were relatively low in all months, there is better convergence between TCR auction prices and actual congestion prices

[25] Notice that these missing months correspond to the missing monthly auction values in Figure 54.

than for the other CSCs. As expected, the TCR auction prices were higher in the summer months than in most other months. However, the highest prices occurred in April and September, following months exhibiting the highest levels of real-time congestion (March and August).

The South to Houston interface showed a more consistent pattern of prices in the monthly TCR auctions than in real-time congestion levels, which varied widely from month to month. The prices paid in the monthly TCR auction were as low as $1 per MWh in the winter, rising to as high as $6 per MWh in June 2003. However, some of the highest congestion prices occurred in the months when the auction prices were low (i.e., March, May, and October). There was a similar pattern on the South-North interface, with a spike in congestion prices in March that far exceeded auction prices.

Normally, one would expect more congestion in the summer months when the load is the highest. However, other factors can cause congestion to be sizable in other months, including transmission outages and sharp fluctuations in fuel costs (e.g., the tight gas conditions that occurred in late February and early March).

To evaluate the total revenue implications of the issues described above, our next analysis compares the TCR auction revenues and obligations. Auction revenues are paid to loads on a load-ratio share basis. Market participants acquire TCRs in the ERCOT-run TCR auction market in exchange for the right to receive TCR credit payments (equal to the congestion price for a CSC times the amount of the TCR). If TCR holders could perfectly forecast shadow prices in the balancing energy market, auction revenues would equal credit payments to TCR holders. The credit payments to the TCR holders should be funded primarily from congestion rent collected in the real-time market from participants scheduling transfers between zones or power flows resulting from the balancing energy market.

The congestion rent from the balancing energy market is associated with the schedules and balancing deployments that result in interzonal transfers during constrained intervals (when there are price differences between the zones). For instance, suppose the balancing energy market deployments result in exports of 600 MWh from the West Zone to the North Zone when the price in the West Zone is $40/MWh and the price in the North Zone is $55/MWh. The

customers in the North Zone will pay $3,300 (600 MWh * $55/MWh) while suppliers in the West Zone will receive $2,400 (600 MWh * $40/MWh). The net result is that ERCOT collects $900 in congestion rent ($3,300 – $2,400) and uses it to fund payments to holders of TCRs. If the quantity of TCRs perfectly matches the capability of the CSC in the balancing energy market, the congestion rent will perfectly equal the amount paid to the holders of TCRs.

Figure 57 reviews the results of these processes by showing (a) monthly and annual revenues from the TCR auctions, (b) credit payments earned by the holders of TCRs based on real-time outcomes, and (c) congestion rent from schedules and deployments in the balancing energy market.

Figure 57: TCR Auction Revenues, Credit Payments, and Congestion Rent 2002 v 2003

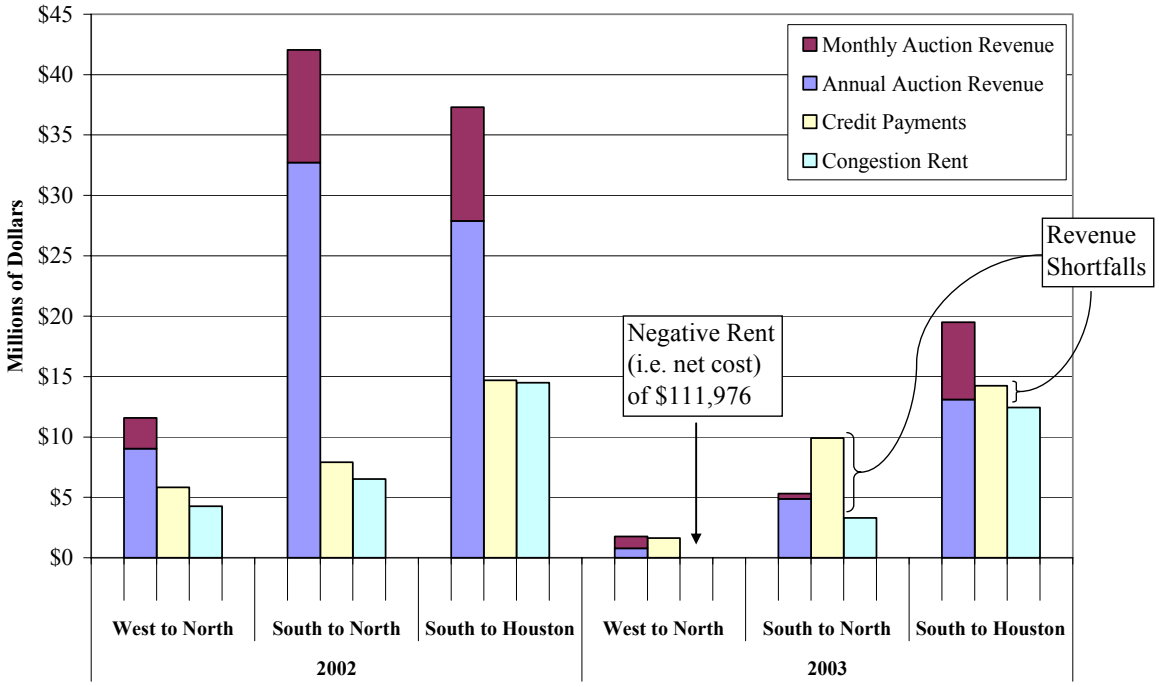


Figure 57 shows that in 2002, the total auction revenues were far greater than credit payments to TCR holders. This is the result of the auction prices being much greater than the average shadow prices that occurred in the balancing energy market (as was shown in Figure 56 above). The figure also shows that from 2002 to 2003, there was a significant reduction in auction revenues (a reduction of 71 percent). Auction revenues were reduced in 2003 because both annual and

monthly auction prices decreased significantly due to improvements in the ability of market participants to forecast congestion on CSCs.

While credit payments for the West to North interface declined by 72 percent between 2002 and 2003, credit payments for the South to North interface increased 25 percent and remained virtually the same for the South to Houston interface. These changes are due primarily to changes in the congestion prices for each CSC. Changes in the quantity of rights sold were also significant.

Figure 57 also shows that the trend in congestion rents is not consistent with the credit payment trend. Congestion rents have decreased from 2002 to 2003 on all three interfaces. Congestion rents from the West to North interface declined from over $4 million to *negative* $111,976, while the rents over the South to North interface declined 49 percent and rents over the South to Houston CSC declined 14 percent. These changes in the congestion rent are due to changes in the congestion price (price difference between zones) as well as changes in the SPD-calculated flow during the constrained intervals.

As described above, a revenue shortfall exists when the credit payments to congestion rights holders exceed the congestion rent. This shortfall is caused when the quantity of congestion rights exceeds the SPD-calculated flow limits in real-time.[26] Revenue shortfalls were only 7 percent of credit payments in 2002, but increased to 39 percent in 2003. These shortfalls are included in the Balancing Energy Neutrality Adjustment charge and assessed to load EROCT-wide. Collecting substantial portions of the congestion costs for the market through such uplift charges reduces the transparency and efficiency of the market. It also increases the costs of transacting and serving load in ERCOT because uplift costs cannot be hedged.

D. Local Congestion and Local Capacity Requirements

In this subsection, we address local congestion and local reliability requirements by evaluating how ERCOT manages the dispatch and commitment of generators when constraints and

[26] For instance, if the shadow price on a particular CSC is $10 per MWh for one hour and the SPD flow limit is 300 MW, ERCOT will collect $3,000 in congestion rents. However, if the holders of congestion rights own a total of 800 MW, then ERCOT must pay out $8,000 worth of credit payments. Thus, the revenue shortfall for ERCOT would be $5,000.

reliability requirements arise that are not recognized or satisfied by the current zonal markets. Local (or intrazonal) congestion occurs in ERCOT when a transmission constraint is binding that is not defined as part of a CSC. Hence, these constraints are not managed by the zonal market model. ERCOT manages local congestion by requesting that generating units adjust their output quantities (either up or down). When not enough capacity is committed to meet reliability, then ERCOT commits additional resources to provide the necessary capacity in either the day-ahead or real-time. Some of this capacity is instructed to be online through Reliability Must Run ("RMR") contracts.

As discussed above, when a unit's dispatch level is adjusted to resolve local congestion, the unit has provided out-of-merit energy or OOME. For the purposes of this report, we define OOME to include both Local Balancing Energy ("LBE") deployed by SPD and OOME deployments, both of which are used to manage local congestion and generally subject to the same settlement rules. Since the output of a unit may be increased or decreased to manage a constraint, the unit may receive an OOME up or an OOME down instruction from ERCOT. Also as explained above, a unit that ERCOT commits to meet its reliability requirements is an out-of-merit commitment or OOMC. The payments made by ERCOT when it takes OOME, OOMC, or RMR actions are recovered through uplift charges to the loads. The payments for each class of action are described below.

When a unit is dispatched out of merit (OOME up or OOME down), the unit is paid for a quantity equal to the difference between the scheduled output based on the unit's resource plan and the OOME instruction from ERCOT. The payment per MWh for OOME is a pre-determined amount specified in the ERCOT Protocols based on the type and size of the unit, the natural gas price, and the balancing energy price. The net payment to a resource receiving an OOME up instruction is equal to the difference between the formula-based OOME up amount and the balancing energy price. For example, for a resource with an OOME up payment amount of $60 per MWh that receives an OOME up instruction when the balancing energy price is $35 per MWh will receive an OOME up payment of $25 per MWh ($60-$35).

For OOME down, the Protocols establish an avoided cost level based on generation type that determines the OOME down payment obligation to the participant. If a unit with an avoided cost

under the Protocols of $15 per MWh receives an OOME down instruction when the balancing energy price is $35 per MWh, then ERCOT will make an OOME down payment of $20 per MWh.

Until August 1, 2003, LBE deployments by SPD were made based on resource-specific bid premiums submitted by the scheduling QSEs. When a "market solution" existed, payments corresponding to the resource specific premiums were made to the QSEs whose resources were selected. A market solution existed when three or more non-affiliated resources were available that could relieve a local transmission constraint. When a market solution does not exist, the LBE is compensated in the same manner as OOME. This process frequently produced results that were not competitive during periods when the market solution test was satisfied. As a result, the market was suspended after August 1, 2003. For the purposes of this report, LBE Up and LBE Down costs and deployments are included with the OOME-up and OOME-down costs and deployments to represent total local congestion management payments and deployments.

A unit providing capacity under an OOMC instruction is paid a pre-determined amount, defined in the ERCOT Protocols, based on the type and size of the unit, natural gas prices, the duration of commitment, and whether the unit incurred start-up costs. Owners of a resource receiving an OOMC instruction from ERCOT are obligated to offer any available energy from the resource into the balancing energy market.

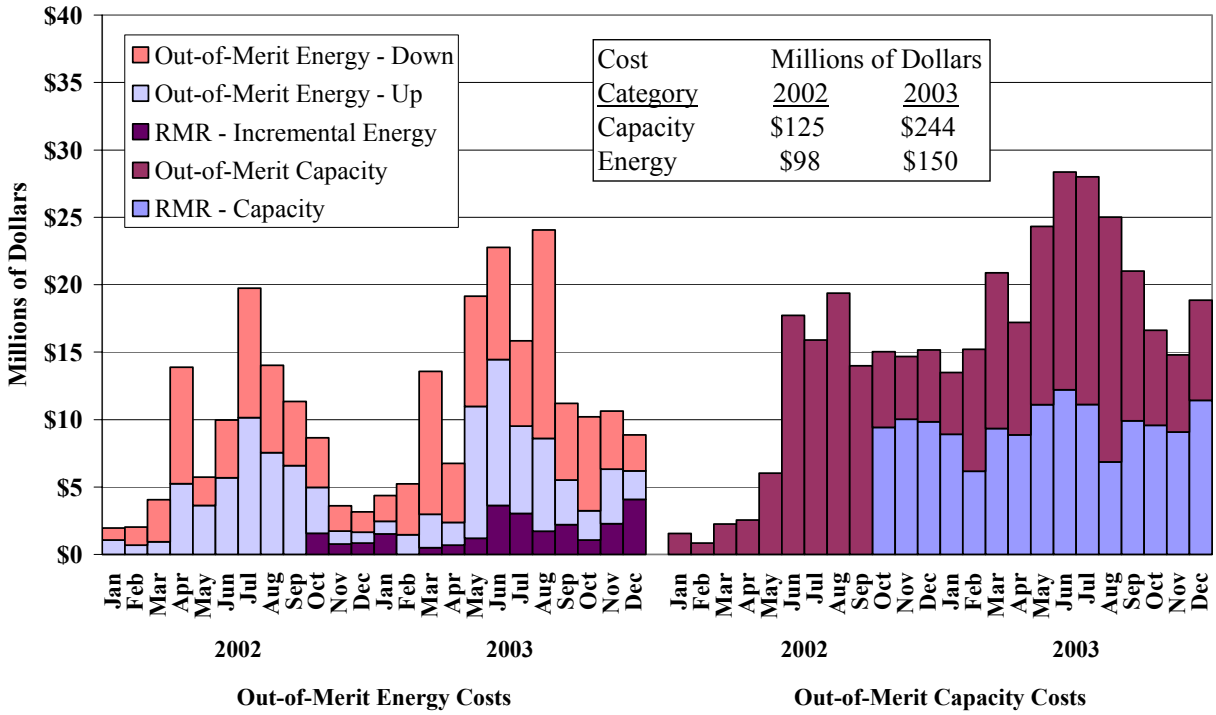
Finally, RMR units committed or dispatched pursuant to their RMR agreements receive cost-based compensation. There were no RMR contracts in ERCOT prior to October of 2002. In response to AEP's announcement that they would place out-of-service all of its gas fired plants in ERCOT because it could buy power at a lower cost than operating the plants, ERCOT contracted with AEP for seven plants to provide RMR service beginning in October 2002. One unit at the Frontera plant in the Rio Grand Valley was also contracted to provide RMR service. Units contracted to provide RMR service to ERCOT are compensated for start-up costs, energy costs, and are also paid a standby fee.[27] The analyses in this section separate RMR uplift into two categories: (a) capacity costs, which include start-up costs, standby fees, and energy costs up

[27] PUCT-MOD, *2002 Annual Report on the ERCOT Wholesale Market*, July 2003, pp. 56-57.

to the minimum dispatch level, and (b) incremental energy costs, which are the costs associated with output above the minimum dispatch level.

Figure 58 shows each of the five categories of uplift costs by month for 2002 and 2003. The left side shows costs of OOME (up and down) and incremental energy from RMR units, while the right side shows the net costs of RMR units and OOMC units. Net cost for RMR units includes only the portion of RMR payments that exceeds the value of energy produced from RMR units at the balancing energy price.

Figure 58: Expenses for Out-of-Merit Capacity and Energy 2002-2003



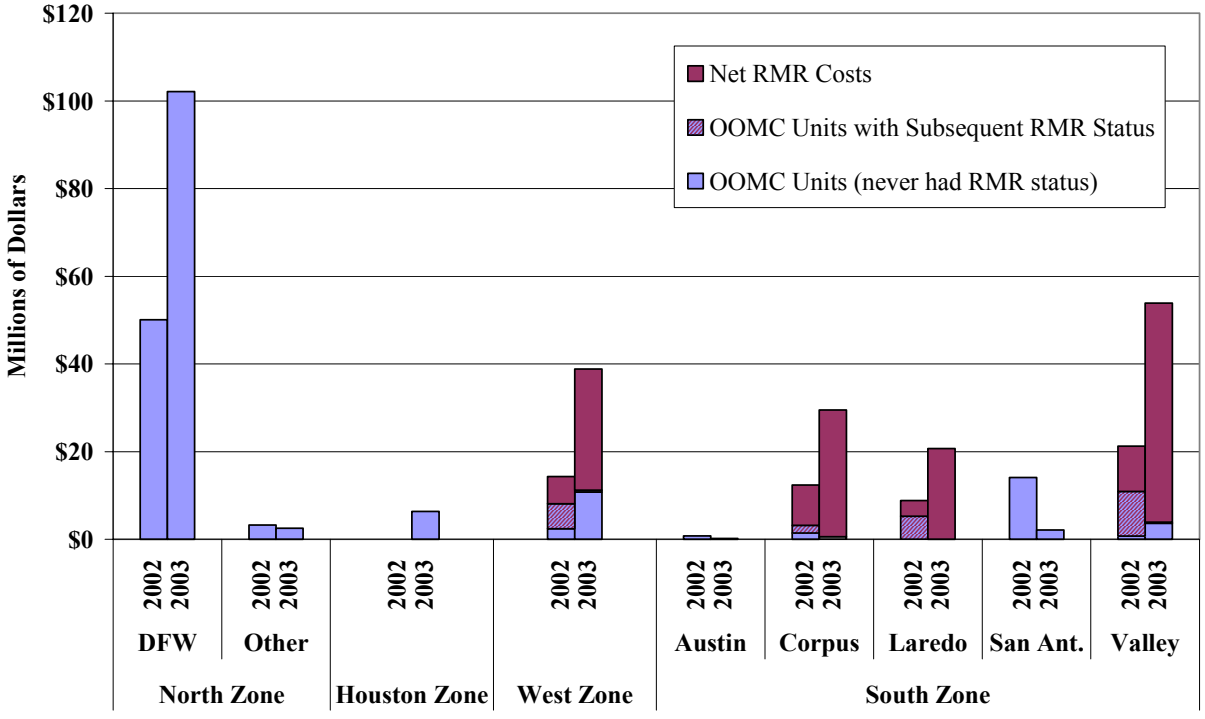
The figure shows that OOME costs and incremental energy costs from RMR units rose from $98 million to $150 million from 2002 to 2003, an increase of 53 percent. Likewise, the costs of OOMC and the capacity costs from RMR units rose 95 percent in 2003. The most substantial increase in uplift costs between 2002 and 2003 was associated with payments to RMR units, which accounted for more than one-third of all out-of-merit uplift costs in 2003. The RMR arrangements were first implemented in October 2002 and, since then, have accounted for a significant portion of uplift costs in each month. The vast majority of the RMR costs are related to recovery of commitment costs and capital costs rather than incremental energy costs.

Out-of-merit costs are greater during the summer when higher loads increase the need for ERCOT operators to take out-of-merit actions to manage local congestion and reliability needs. However, RMR costs do not increase substantially during the summer months because RMR payments are primarily designed to recover fixed costs, which are constant throughout the year.

Although the costs are borne by load throughout ERCOT, the costs are caused in specific locations because most of these actions are taken to maintain local reliability. The rest of the analyses in this section evaluate in more detail where these costs were caused and how they have changed from 2002 to 2003. The first of these analyses focuses on the payments made for commitment of capacity, which include OOMC and RMR payments. Figure 59 shows these payments by location for 2002 and 2003.

OOMC costs in this figure are divided into two groups of units: (a) units that switched to RMR status after October 2002, and (b) units that never switched to RMR status. For units that became RMR units during 2002, the OOMC costs prior to October 2002 are shown separately in the shaded areas. This is done to allow a more direct comparison of the RMR capacity costs in 2003 to the commitment costs (RMR commitment and OOMC) for the same RMR units in 2002.

Figure 59: Expenses for OOMC and RMR by Region 2002-2003



The total commitment-related uplift costs, including OOMC and RMR payments for capacity costs, increased from $125 million in 2002 to $244 million in 2003, an increase of nearly 100 percent. Roughly half of these costs in 2003 are payments to units providing OOMC. The largest source of OOMC uplift costs is the Dallas/Fort Worth area, accounting for 52 percent of the OOMC costs in 2002 and 79 percent in 2003. No units in this area or other areas within the North Zone or Houston Zones were designated as RMR resources. OOMC costs in the West Zone increased by almost 39 percent in 2003, despite the fact that a number of the units that had received OOMC payments in 2002 were subsequently designated as RMR units.

OOMC costs in the South Zone decreased from $34 million in 2002 to slightly less than $7 million in 2003. Most of this decrease is due to the fact that most of the resources that received OOMC payments in 2002 were designated as RMR resources and received RMR payments in 2003. To make a more direct comparison of OOMC costs between 2002 and 2003, it is useful to compare the OOMC costs in the South Zone and West Zone for units that were never designated as RMR resources. Controlling for this, OOMC costs for the South Zone and West Zone increased from $19 million in 2002 to $26 million in 2003.

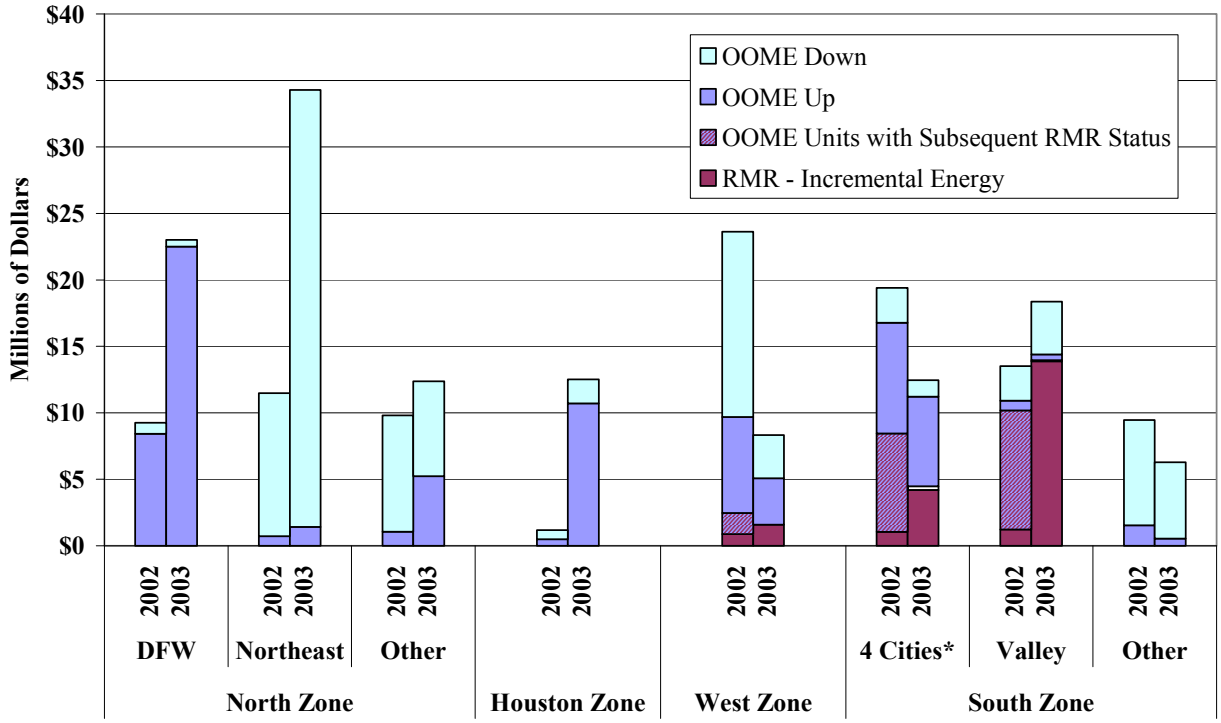
The largest share of the increased commitment costs in 2003 is associated with units that have been designated as RMR resources. The payments to these units increased from $52 million in 2002 (including OOMC payments prior to RMR designation) to $128 million in 2003. Since October 2002, the largest source of RMR capacity uplift costs has been the South Zone with 78 percent of the RMR costs. The remaining 22 percent of the RMR capacity costs are incurred on units in the West Zone.

Much of the increases in the RMR and OOMC costs in 2003 can be attributed to the increase in natural gas prices in 2003. Both the RMR and OOMC payments are derived from formulas that index a share of the payments to natural gas prices. Additionally, the increase in payments to units that have been designated as RMR resources is not surprising because the OOMC payments that these resources received in 2002 would not have explicitly included recovery of long-term fixed costs as the RMR payments do.

The next analysis reviews the costs incurred by ERCOT to dispatch generating resources out of merit to resolve local congestion. The costs are incurred in the form of OOME up and OOME

down payments, as well as payments to RMR resources for incremental energy above minimum generation.[28] Figure 60 shows annual uplift costs for units providing OOME by region and by zone.

Figure 60: Expenses for OOME by Region
2002-2003



Note: There was an additional $22 million in uplift costs in 2003 for which the specific location of the generator was not identified. Of this amount, $18.4 million was paid for OOME down service in the North Zone and $3.0 million was paid for OOME down service in the South Zone.

This figure shows that uplift for OOME increased from $31 million in 2002 to $88 million in 2003 in the North zone. It increased from $1 million in 2002 to $13 million in 2003 in Houston and decreased from $23 million in 2002 to $7 million in 2003 in the West Zone. In the South Zone, uplift decreased from $40 million in 2002 to $22 million in 2003.

Uplift is shown for three areas within the North Zone: (a) Dallas/Fort Worth, (b) the area which became the Northeast Zone in 2004, and (c) all other areas in the North Zone. Dallas/Fort Worth is a load pocket and so units within it are asked to dispatch upwards to relieve congestion, however, the Northeast is a generation pocket where units are dispatched down to relieve

[28] Local balancing energy is included in the OOME costs as described above.

congestion. Before 2004, these areas were within the same pricing region and so there was no market-based mechanism for dispatching more energy in Dallas/Fort Worth while dispatching less in the Northeast. The addition of the Northeast Zone in 2004 should lead to substantial reductions in OOME down dispatch.

Uplift costs decreased substantially in West Zone from 2002 to 2003 for two primary reasons. First, several units switched to RMR status and this has led to some reclassification of OOME costs to RMR costs. Second, the compensation rules for wind resources changed, resulting in decreased uplift payments to wind resources.

Uplift costs decreased substantially in the South Zone from 2002 to 2003 at least partly because a number of units switched to RMR status resulting in a reclassification of their costs. The South Zone, like the North, is a vast area containing both generation pockets and load pockets. Generally, the populous areas of Austin, San Antonio, Laredo, and Corpus Christi are more likely to be dispatched up out of merit while the outlying areas are more likely to be dispatched down out of merit. Like the break-up of the North into separate zones, splitting up the South Zone into multiple zones might reduce uplift costs for local transmission congestion.

E. Conclusions and Recommendations: Interzonal and Intrazonal Congestion

The results in this section highlight significant opportunities for improvements in the operation of the ERCOT markets. These results indicate that in 2003, the vast majority of the congestion costs are associated with intrazonal congestion. This process results in uplift that is difficult to hedge and that is inefficiently allocated to the load in ERCOT. The process also results in economic signals that are not transparent. In addition, the intrazonal congestion management procedures appear to provide incentives for some suppliers to submit inaccurate resource plans to increase the frequency of out of merit commitment and dispatch actions by ERCOT

With regard to interzonal congestion, the report highlights significant issues related to the zonal assumptions used in the ERCOT market. These assumptions and the operation of the current markets in Texas will be further evaluated in an operations report to be issued this fall. However this report indicates that:

- The current zonal market can result in large inconsistencies between the interzonal flows calculated by SPD and the actual flows over the CSC interfaces; and
- These inconsistencies can result in under-utilized transmission capability and difficulties in defining transmission rights whose obligations can be fully satisfied.

The most complete long-run remedy for both the interzonal and intrazonal issues identified in this report would be to implement nodal markets, an option that is currently being evaluated in ERCOT. These markets would provide transparent prices for both generators and loads that would fully reflect all transmission constraints on the ERCOT network. Hence, we strongly recommend the continued development and implementation of such markets.

In the short-term, this report recommends additional changes to address intrazonal congestion that are also discussed in Section II. We recommend the creation of a zone for Dallas/Ft. Worth. This would allow a large share of the congestion that is currently managed with OOME processes to be priced more efficiently and transparently. We understand that there would be significant issues to consider in forming such a zone, including the effect on current bilateral contracts, the need for measures to effectively mitigate market power in the area, and the equity implications of such a change. In addition, the benefits of defining a new zone are based on the assumption that CSCs between Dallas-Fort Worth and adjacent areas could be defined that include the key transmission constraints that currently result in OOME and OOMC actions by ERCOT. This would need to be analyzed and validated by ERCOT.

V. ANALYSIS OF COMPETITIVE PERFORMANCE

In this section, we evaluate competition in the ERCOT market by analyzing the market structure and the conduct of the participants during 2003.

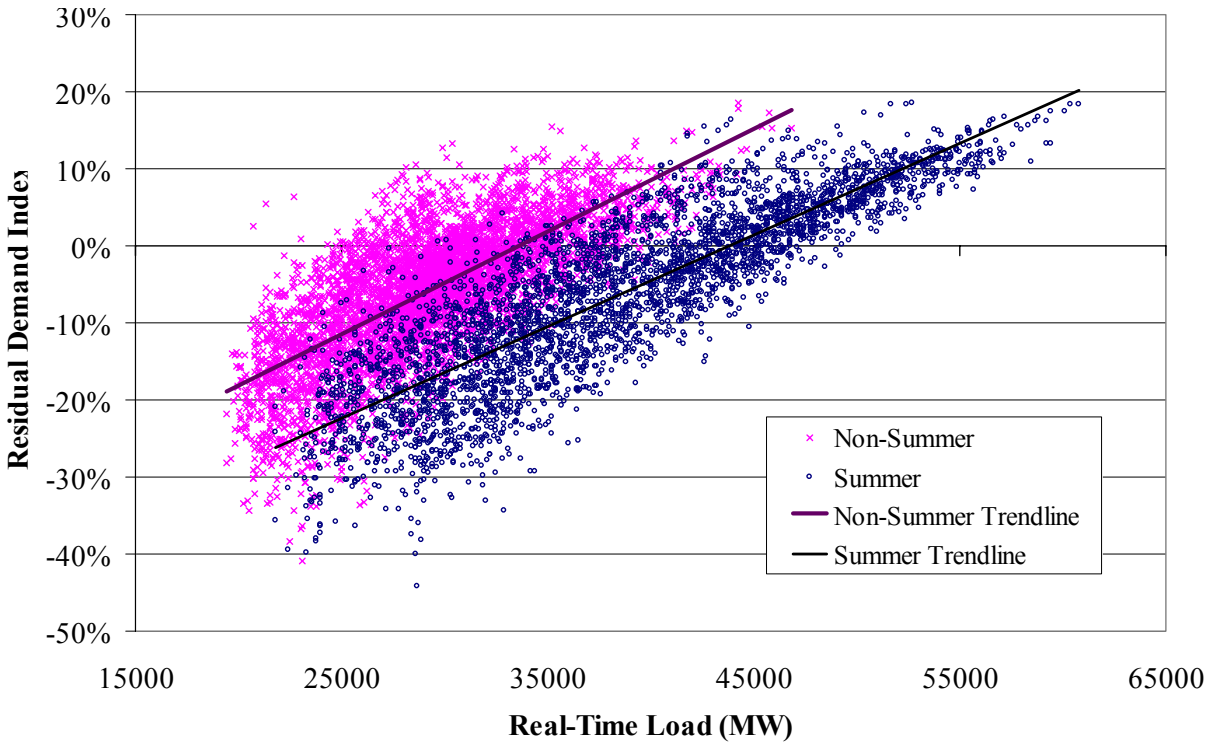
A. Structural Market Power Indicators

We analyze market structure using the Residual Demand Index ("RDI"), a statistic that measures the percentage of load that could not be satisfied without the resources of the largest supplier. When the RDI is greater than zero, the largest supplier is pivotal (i.e. its resources are needed to satisfy the market demand). When the RDI is less than zero, no single supplier's resources are required in order to serve the load as long as the resources of its competitors are available.

The RDI is a useful structural indicator of potential market power, although it is important to recognize its limitations. As a structural indicator, it does not illuminate actual supplier behavior, indicating whether a supplier may have exercised market power. The RDI also does not indicate whether it would be profitable for a pivotal supplier to exercise market power. However, it does identify conditions under which a supplier would have the *ability* to raise prices significantly by withholding resources.

Figure 61 shows the RDI for both summer hours and non-summer hours relative to load. The trend lines for each data series are also shown and indicate a strong positive relationship between load and the RDI. This relationship is expected since the quantity of resources available from competing suppliers would have to increase as load increases to keep the RDI from increasing.

Figure 61: Residual Demand Index
Summer and Non-Summer Hours - 2003



The figure shows that the RDI in the summer generally begins to be positive in many hours when load exceeds 35,000 MW. During the entire summer, the RDI was greater than zero in more than 1,100 hours. During other periods, the RDI is frequently positive when the load rises above 25,000 MW. The RDI can be positive at lower load levels during other periods due to large quantities of generation taking planned outages in the spring and fall. Hence, although the load is lower in these periods, our analysis shows that a supplier is pivotal in more than 1,200 hours. The effects of the planned outages in non-summer periods can also be seen by the difference in the trend lines for the two periods, namely, the trend line for the non-summer hours is 14 percent higher than in the summer hours.

It is important to recognize that inferences regarding market power cannot be made solely from this data. Some of the largest suppliers also serve substantial load, which causes them to be a much smaller *net* seller than the analysis above would indicate. For example, a smaller supplier selling energy in the balancing energy market and through short-term bilateral contracts may have a much greater incentive to exercise market power than a larger supplier with long-term

contracts and load obligations. To account for this factor, we also calculated a load-adjusted RDI.

The "load-adjusted" RDI is adjusted for the load served by each supplier. Thus, a supplier with 3,000 MW of capacity and 2,000 MW of load would have a "load-adjusted capacity" of 1,000 MW and only the load-adjusted capacity is used in calculating the RDI. The logic underlying this alternative method is that the supplier would not have the incentive to withhold more than 1,000 MW because it would have to purchase that additional amount from the balancing market to serve its load (assuming that the supplier has no other physical or financial contracts to purchase energy, which may or may not be the case). Because many suppliers may have substantial contractual positions and because some of the load may be served on a relatively short-term basis, the true RDI for the largest suppliers is likely to lie between unadjusted values shown in Figure 61 and the load-adjusted RDI values shown in Figure 62. Figure 62 shows the load-adjusted RDI for ERCOT as a function of the actual load level for both summer and non-summer hours.

Figure 62: Load-Adjusted Residual Demand Index vs. Actual Load Summer and Non-Summer Hours -- 2003

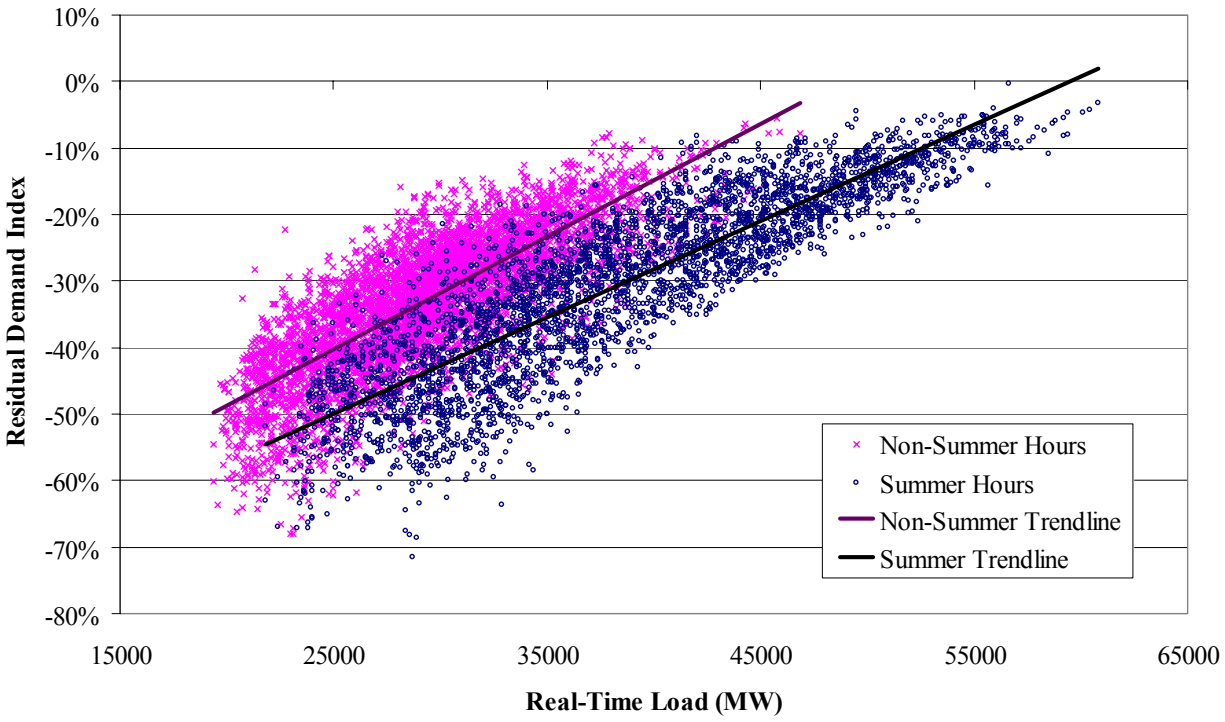
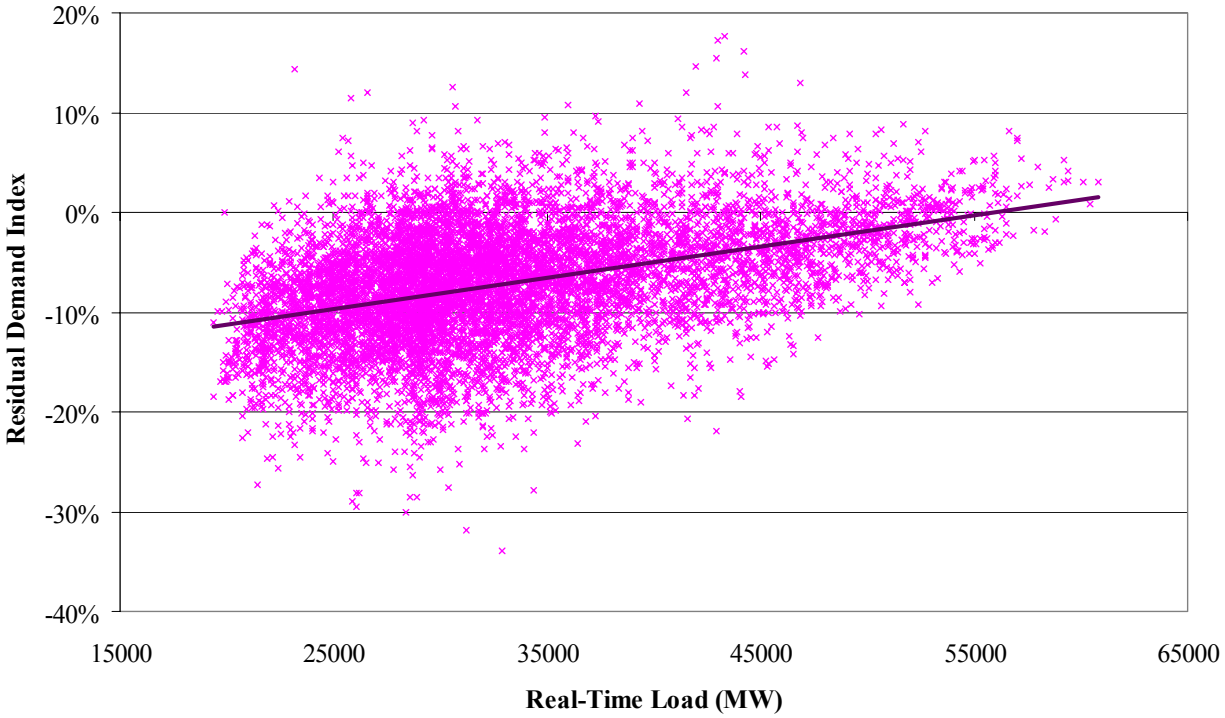


Figure 62 shows that there were no hours in the summer or non-summer periods with a positive RDI, although some hours were very close. Hence, no suppliers were pivotal by this measure. This RDI measure does not consider the contractual position of the supplier, which can increase a supplier's incentive to exercise market power compared to the load-adjusted capacity assumption made in this analysis. The PUCT is now collecting bilateral contract information that could potentially be used to improve the accuracy of this measure.

In addition, a supplier's ability to exercise market power in the current ERCOT balancing market will generally be higher than indicated by the load-adjusted RDI because a significant share of the available energy resources in real time are not offered in the ERCOT balancing market (as shown in prior sections of this report). Hence, a supplier may be pivotal in the balancing energy market when it would not have been pivotal more generally. To account for this, we developed RDI statistics for the balancing energy market. Figure 63 shows the RDI in the balancing energy market relative to the actual load level.

Figure 63: Balancing Energy Market Residual Demand Index vs. Actual Load 2003



Ordinarily, the RDI is used to measure the percentage of load that cannot be served without the resources of the largest supplier, assuming that the market could call upon all committed and quick-start capacity owned by other suppliers. Figure 63 limits the other supplier's capacity to the energy offered in the balancing energy market. When the RDI is greater than zero, the largest supplier's balancing energy offers are necessary to prevent a price spike in the balancing energy market.

While the RDI was negative in the majority of hours, it was positive in 11 percent of all hours and in 24 percent of hours when real-time load exceeded 40 GW. The instances when the RDI was positive occurred over a wide range of load levels, from 25 GW to 60 GW. The RDI results for the balancing energy market shown in Figure 63 help explain how transient price spikes can occur under mild demand while large amounts of capacity are available in ERCOT. These results also show how current market issues that cause QSEs to offer only part of their available energy in the balancing energy market can cause the balancing energy market to be vulnerable to withholding and other forms of market abuses even when no suppliers are fundamentally pivotal (i.e., the load-adjusted RDI is positive). This highlights the importance of modifying the current market rules and procedures to minimize any barriers or disincentives to full participation in the balancing energy market.

B. Evaluation of Supplier Conduct

The prior section is a structural analysis in ERCOT that allows one to draw inferences about potential market power. This section evaluates actual participant conduct to identify evidence of attempts to exercise market power through physical and economic withholding. In particular, we examined unit deratings and forced outages to detect physical withholding and we evaluate the "output gap" to detect economic withholding.

In a single-price auction like the balancing energy market auction, suppliers may attempt to exercise market power by withholding resources. The purpose of withholding is to cause more expensive resources to set higher market clearing prices, allowing the supplier to profit on its other sales in the balancing energy market. Because forward prices will generally be highly correlated with spot prices, price increases in the balancing energy market can increase a

supplier's profits in the bilateral energy market. The strategy is profitable when the withholding firm's incremental profit is greater than the lost profit from the sales of its withheld capacity.

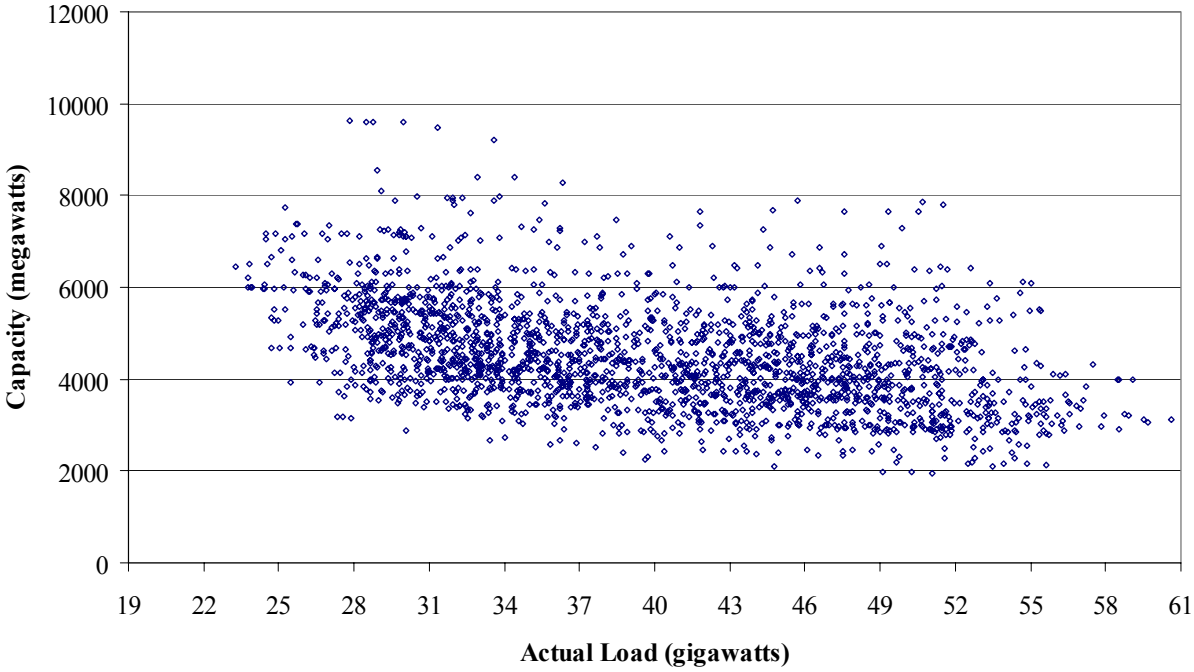
1. Evaluation of Potential Physical Withholding

Physical withholding occurs when a participant makes resources unavailable for dispatch that are otherwise physically capable of providing energy and which are economic at prevailing market prices. This is done by derating the unit or designating it as a forced outage. In any electricity market, deratings and forced outages are unavoidable. The goal of the analysis in this section is to differentiate these justifiable deratings and outages from physical withholding. We test for physical withholding by examining deratings and forced outage data to ascertain whether the data is correlated with conditions under which physical withholding would likely be most profitable.

The RDI results shown in Figure 61 and Figure 62 indicate that the potential for market power abuses rises as load rises and RDI values become more positive. Hence, if physical withholding is a persistent problem in ERCOT, we would expect to see increased deratings and forced outages at the highest load levels. Alternatively, because competitive prices increase as load increases, deratings and forced outages in a market performing competitively will tend to decrease as load approaches peak levels. Suppliers that lack market power will take actions to maximize the availability of their resources since their output is generally most profitable in these peak periods.

Figure 64 shows the relationship of short-term deratings and forced outages to real-time load levels in each hour during the summer months. We focus on these months to eliminate the effects of planned outages and other discretionary deratings that occur in off-peak periods. Long-term deratings are not included in this analysis because they are far less likely to constitute physical withholding given the costs of such withholding. Renewable resources and cogeneration resources are also excluded from this analysis given the high variation in the availability of these classes of resources.

Figure 64: Short-Term Deratings and Forced Outages vs. Actual Load
June to August, 2003

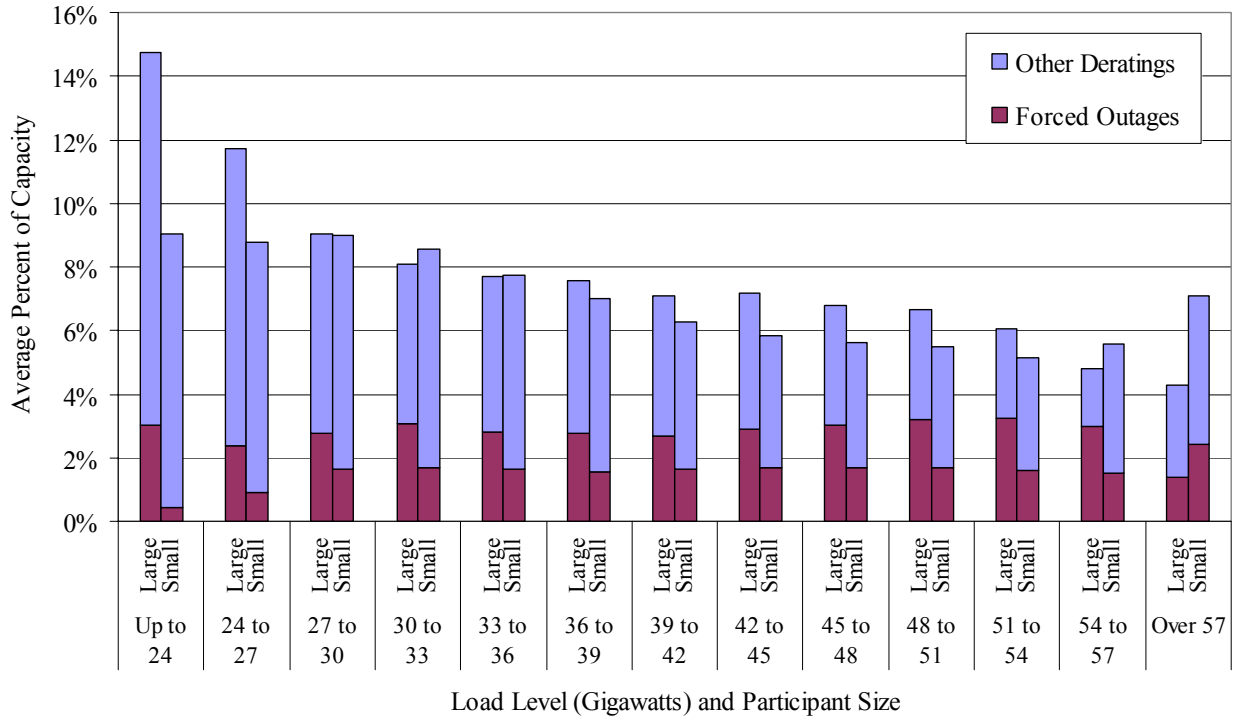


Short-term deratings and outages varied between 2 GW and 10 GW from June to August 2003. Figure 64 reveals a slight downward trend between deratings and outages and real-time demand levels. At demand levels above 56 GW, the sum of deratings and outages were generally near or less than 4 GW. This is notable because at the highest demand levels, resources that are seldom dispatched and generally less reliable must be called on to satisfy the market's energy requirements. The results in Figure 64 are consistent with the conclusion that most suppliers have competitive incentives to increase their resource availability under peak demand conditions when energy sales are most profitable.

However, we further evaluate these trends by examining them by portfolio size. Portfolio size is important in determining whether individual suppliers have incentives to withhold available resources. Hence, the patterns of outages and deratings of large suppliers can be usefully evaluated by comparing them to the small suppliers' patterns. Figure 65 shows the average relationship of short-term deratings and forced outages as a percentage of total installed capacity

to real-time load level during the summer months for large and small suppliers.[29] The large supplier category includes the four largest suppliers in ERCOT, whereas the small supplier category includes the remaining suppliers (as long as the supplier controls at least 300 MW of capacity).[30]

Figure 65: Short-Term Deratings by Load Level and Participant Size
June to August, 2003



For large suppliers, the short-term derating or forced outage rates decreased from approximately 15 percent at low demand levels to about 4 percent at load levels above 57 GW. For small suppliers, the derating rates decreased from 9 percent at load levels below 24 GW to less than 6 percent at load levels between 54 GW and 57 GW. The deratings and outages for small suppliers rose to more than 7 percent in the small number of hours when demand exceeded 57 GW.

At most load levels, large and small suppliers have comparable derating rates. However, large suppliers have larger levels of deratings and outages at low load levels and small suppliers have

29 Like the prior analysis, long-term deratings and deratings by cogeneration and renewable energy resources are excluded.

30 The four largest suppliers are Texas Utilities, Texas Genco, AEP, and Calpine.

higher portions at the highest load levels. Given that the market is most vulnerable to market power at the highest load levels, these derating patterns do not show evidence of physical withholding by the large suppliers. However, these results cannot exclude limited instances of withholding by either large or small suppliers. Such instances can only be identified through a detailed investigation.

Figure 65 also shows that a larger share of the large suppliers' deratings was comprised of forced outages. Given the extremely low forced-outage rates shown for small suppliers, it is likely that this difference is due, in part, to differences in forced outage reporting by smaller suppliers. The fact that the total deratings of the small suppliers rises substantially at the highest load conditions could indicate a potential concern that should be further investigated by MOD.

2. Evaluation of Potential Economic Withholding

To complement the prior analysis of physical withholding, this subsection evaluates potential economic withholding by calculating an "output gap". The output gap is defined as the quantity of energy that is not being produced by in-service capacity even though the in-service capacity is economic by a substantial margin given the balancing energy price. A participant can economically withhold resources, as measured by the output gap, by raising the balancing energy offers so as not to be dispatched (including both balancing up and balancing down offers) or by not offering unscheduled energy in the balancing energy market.

Resources can be included in the output gap when they are committed and producing at less than full output or when they are uncommitted and producing no energy. Unscheduled energy from committed resources is included in the output gap if the balancing energy price exceeds the marginal production cost of the energy by at least $50 per MWh. Uncommitted capacity is considered to be in the output gap if the unit would have been substantially profitable given the prevailing balancing energy prices. The resource is counted in the output gap if its net revenue (market revenues less incremental production costs) exceeds the minimum commitment costs of the resource (including start-up and no-load costs) by a margin of at least $50 per MWh for its minimum output level over its minimum run-time.[31]

[31] The production costs are estimated using the Continuous Emissions Monitoring ("CEMS") data collected by the Environmental Protection Agency. This data is used to estimate incremental heat rates and heat

Like the outages and deratings, the output gap will frequently detect conduct that can be competitively justified. Hence, it is important to evaluate the correlation of the output gap patterns to those factors that increase the potential for market power, including load levels and portfolio size. Figure 66 shows the relationship between the output gap from committed resources and real-time load for all hours during 2003.

Figure 66: Output Gap from Committed Resources vs. Actual Load 2003

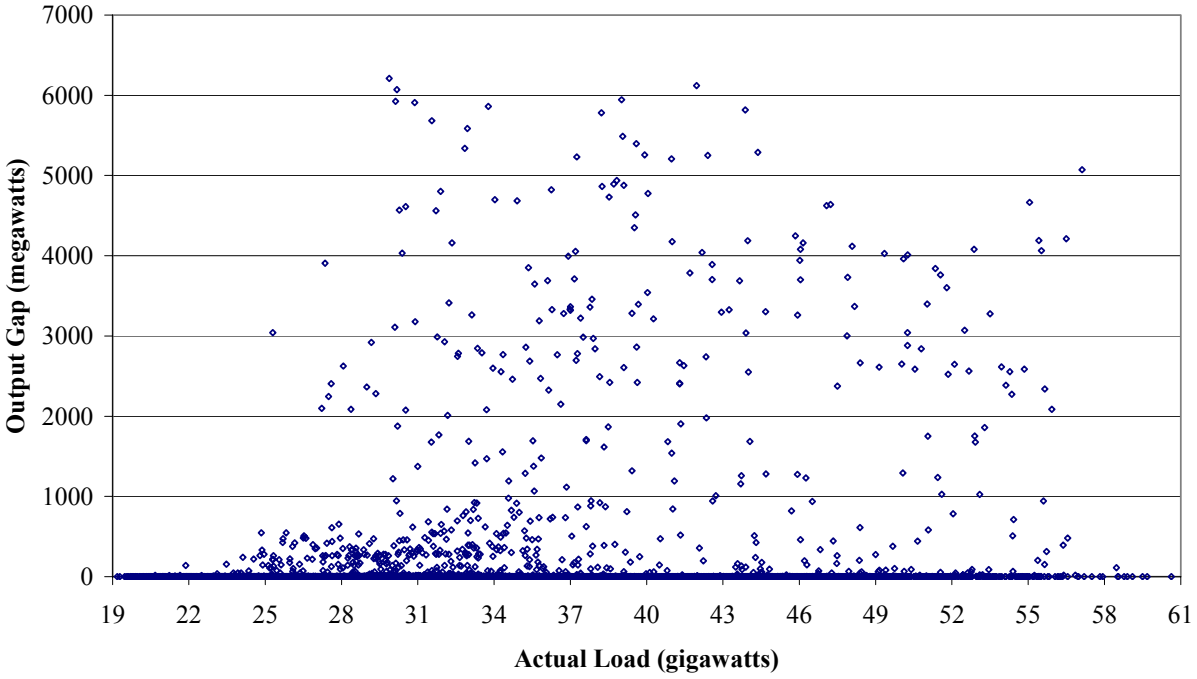


Figure 66 shows that the output gap from committed resources ranged from zero in most hours to a maximum of 6,210 MW during 2003. This figure also shows that there is no clear relationship between the output gap and real-time demand. The high output gap values generally occurred during transitory price spikes that occurred at a wide range of demand levels and tend to make most of the unscheduled energy appear economic. The transitory nature of most of these instances would make a large share of the identified output unavailable due to the resources' ramp limitations. Ramp limitations dictate that resources cannot respond instantaneously to an

input at minimum generation levels for ERCOT generating units. This analysis also assumes $4 per MWh variable operating and maintenance expenses. Whenever CEMS data is unavailable, minimum generation and incremental costs are estimated by looking at a sample of balancing energy prices that coincide with each resource's production over the previous 90 days.

unpredicted price spike. In addition, quick-start resources are frequently unable to respond quickly enough in response to a transitory price spike, particularly if it is unforeseen. The next analysis further examines the output gap results by showing it by size of supplier and load level.

Figure 67 compares real-time load to the average output gap as a percentage of total installed capacity by participant size. The large supplier category includes the four largest suppliers in ERCOT,[32] whereas the small supplier category includes the remaining suppliers that control more than 300 MW of capacity. The output gap is separated into (a) quantities associated with uncommitted resources and (b) quantities associated with incremental output ranges of committed resources.

Figure 67: Output Gap by Load Level and Participant Size
2003

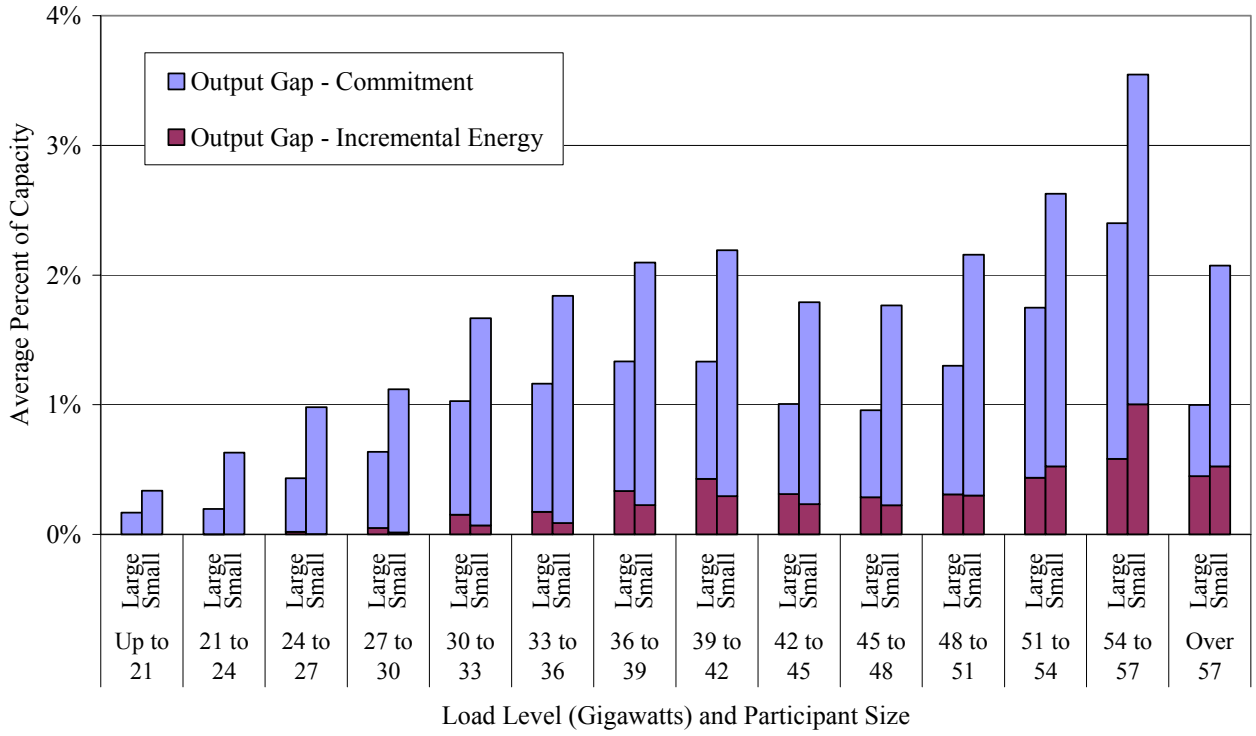


Figure 67 shows that the large suppliers' total output gap was lower at every load level than that of the smaller suppliers. Additionally, the output gap associated with incremental energy on committed resources was lower for large suppliers than the comparable amount for small

[32] The four largest suppliers are Texas Utilities, Texas Genco, AEP, and Calpine.

suppliers. To the extent the conduct of the smaller suppliers serves as a competitive benchmark, these results do not indicate that the large suppliers engaged in economic withholding. However, the figure also shows that the output gap quantities generally increased as load levels increased.

Large suppliers' output gap increased from close to zero at low demand levels to over 2 percent when demand levels were between 54 GW and 57 GW. For small suppliers, the output gap increased in a similar pattern from close to zero at load levels below 24 GW to over 3 percent at load levels between 54 GW and 57 GW. At the very highest load levels, large and small suppliers' output gaps decreased (to approximately 1 percent for large suppliers and to just over 2 percent for small suppliers).

Although the total output gap increased as load increased, we do not conclude that economic withholding has been a significant concern in ERCOT for a number of reasons. First, the output gap will naturally rise as prices rise since a larger share of the generation base will be economic. Hence, the output gap can increase with load (and prices) even when participants' conduct is unchanged.

Second, the output gap associated with incremental energy from large suppliers remains near or below one-half of one percent at all load levels, while the incremental output gap for small suppliers remains less than one percent. Lastly, the largest share of the total output gap is the output associated with uncommitted resources. This output gap category raises fewer concerns than the incremental energy output gap because our prior analysis clearly indicated that the ERCOT market is generally over-committed. Hence, the economic opportunity to profitably commit additional resources would not usually be foreseen by market participants.

The largest output gap quantities are associated with resources owned by small utilities. While we have generally treated the small suppliers' conduct as competitive for the purpose of drawing conclusions about the large suppliers' conduct, it is possible under high load conditions that some of this conduct could represent withholding. We believe it is more likely that these results reflect a desire on the part of many small suppliers to serve their own load and not participate actively in the ERCOT markets. Nevertheless, a review of their conduct in select high-load hours may be warranted.

Based on the analyses in this section of the report, there is no clear indication that suppliers have systematically exercised market power by economically or physically withholding capacity. However, this report is limited to evaluating overall patterns of conduct. Isolated instances of significant physical or economic withholding would generally need to be identified on a case-specific basis.

APPENDIX A

Frequent OOMC Resources

Resource	OOMC Uplift per MWh of Production	QSE	Zone	Subzone
GEN_SILASRAY_SILAS_5	$81.71	BROWNSVILLE PUBLIC UTILITY BOARD TENASKA (RES)	SOUTH2003	Valley
GEN_RIOP_RIOP_G5	$66.91	WEST TEXAS UTILITIES CO	WEST2003	West
GEN_HLSES_UNIT2	$43.79	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_SAPS_SAPS_G1	$41.99	WEST TEXAS UTILITIES CO	WEST2003	West
GEN_EMSES_UNIT3	$34.33	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_HLSES_UNIT4	$33.52	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_SILASRAY_SILAS_9	$32.33	BROWNSVILLE PUBLIC UTILITY BOARD TENASKA (RES)	SOUTH2003	Valley
GEN_EMSES_UNIT1	$32.03	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_HLSES_UNIT5	$25.62	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_EMSES_UNIT2	$24.36	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_SILASRAY_SILAS_6	$23.44	BROWNSVILLE PUBLIC UTILITY BOARD TENASKA (RES)	SOUTH2003	Valley
GEN_MCSES_UNIT6	$21.36	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_MCSES_UNIT7	$19.74	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_FTPP_FTPP_G1	$16.92	WEST TEXAS UTILITIES CO	WEST2003	West
GEN_MGSES_UNIT5	$15.77	TXU ELECTRIC CO (RES)	WEST2003	West
GEN_HLSES_UNIT3	$15.44	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_LHSES_UNIT1	$15.26	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_MCSES_UNIT8	$11.83	TXU ELECTRIC CO (RES)	NORTH2003	DFW

Frequent OOM Up Units

Resource	OOM-Up Uplift per MWh of Production	QSE	Zone	Subzone
GEN_SPNCER_SPNCE_4	$18.62	DENTON MUNICIPAL ELECTRIC (RES)	NORTH2003	DFW
GEN_SPNCER_SPNCE_3	$13.15	DENTON MUNICIPAL ELECTRIC (RES)	NORTH2003	DFW
GEN_SPNCER_SPNCE_1	$13.02	DENTON MUNICIPAL ELECTRIC (RES)	NORTH2003	DFW
GEN_MCSES_UNIT6	$7.42	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_MCSES_UNIT7	$7.38	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_LHSES_UNIT1	$6.33	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_NLSES_UNIT3	$5.72	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_DPW_DPW_G7	$5.31	RELIANT ENERGY HL AND P (RES)	HOUSTON2003	Houston
GEN_MCSES_UNIT8	$3.86	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_SPNCER_SPNCE_5	$3.82	DENTON MUNICIPAL ELECTRIC (RES)	NORTH2003	DFW
GEN_NLSES_UNIT1	$3.50	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_SRB_SRB_G4	$3.40	RELIANT ENERGY HL AND P (RES)	HOUSTON2003	Houston
GEN_LHSES_UNIT2	$3.38	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_HLSES_UNIT4	$2.90	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_HLSES_UNIT5	$2.81	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_EMSES_UNIT3	$2.78	TXU ELECTRIC CO (RES)	NORTH2003	DFW
GEN_SRB_SRB_G3	$2.72	RELIANT ENERGY HL AND P (RES)	HOUSTON2003	Houston

Frequent OOM Down Units

Resource	OOM-Down Uplift per MWh of Production	QSE	Zone	Subzone
GEN_VLSES_UNIT2	$2.31	TXU ELECTRIC CO (RES)	NORTH2003	North-East
GEN_MNSES_UNIT1	$1.77	TXU ELECTRIC CO (RES)	NORTH2003	North-East
GEN_CHE_CHEST1	$1.67	CALPINE CORP	HOUSTON2003	Houston
GEN_MNSES_UNIT3	$1.28	TXU ELECTRIC CO (RES)	NORTH2003	North-East
GEN_MNSES_UNIT2	$1.16	TXU ELECTRIC CO (RES)	NORTH2003	North-East