POTOMAC ECONOMICS

2005 STATE OF THE MARKET REPORT
FOR THE
ERCOT WHOLESALE ELECTRICITY MARKETS

POTOMAC ECONOMICS, LTD.

Advisor to Wholesale Market Oversight
Public Utility Commission of Texas

July 2006

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

We wish to acknowledge the helpful input and numerous comments provided by the staff of the Wholesale Market Oversight of the Public Utility Commission of Texas, including Parviz Adib, Richard Greffe, and Danielle Jaussaud. 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

This report reviews and evaluates the outcomes of the ERCOT wholesale electricity markets in 2005. It includes assessments of the incentives provided by the current market rules and procedures, and analyses of the conduct of market participants. We find improvements in a number of areas over the results in prior years that can be attributed to changes in the market rules or operation of the markets. However, the report generally confirms prior findings that the current market rules and procedures are resulting in systematic inefficiencies.

These findings can be found in three previous reports we have issued regarding the ERCOT electricity markets.[1] These reports included a number of recommendations designed to improve the performance of the current ERCOT markets. Many of these recommendations were considered by ERCOT working groups and some were embodied in protocol revision requests ("PRRs"). Most of the remaining recommendations will be addressed by the introduction of a nodal market design, which is currently being developed for implementation in 2009.

The wholesale market should function more efficiently under the nodal market design by: providing better incentives to market participants, facilitating more efficient commitment and dispatch of generation, and improving ERCOT's operational control of the system. The congestion on all transmission paths and facilities will be competitively managed through the nodal market, which is in contrast to the current zonal market design, where most transmission congestion is resolved through non-transparent, non-market-based procedures.

Under the nodal market, unit-specific dispatch will allow ERCOT to more fully utilize the generating resources than the current market, which frequently exhibits shortage prices when the generating capacity is not fully utilized. Finally, the nodal pricing will result in price signals that provide incentives to build new generation where it is most needed for managing congestion and maintaining reliability. In the long-term, these enhancements to overall market efficiency should translate into substantial savings for consumers.

---

[1] "ERCOT State of the Market Report 2003", Potomac Economics, August 2004 (hereafter "2003 SOM Report"); "2004 Assessment of the Operation of the ERCOT Wholesale Electricity Markets", Potomac Economics, November 2004 (hereafter "Assessment of Operations"); and "ERCOT State of the Market Report 2004", Potomac Economics, July 2004 (hereafter "2004 SOM Report").

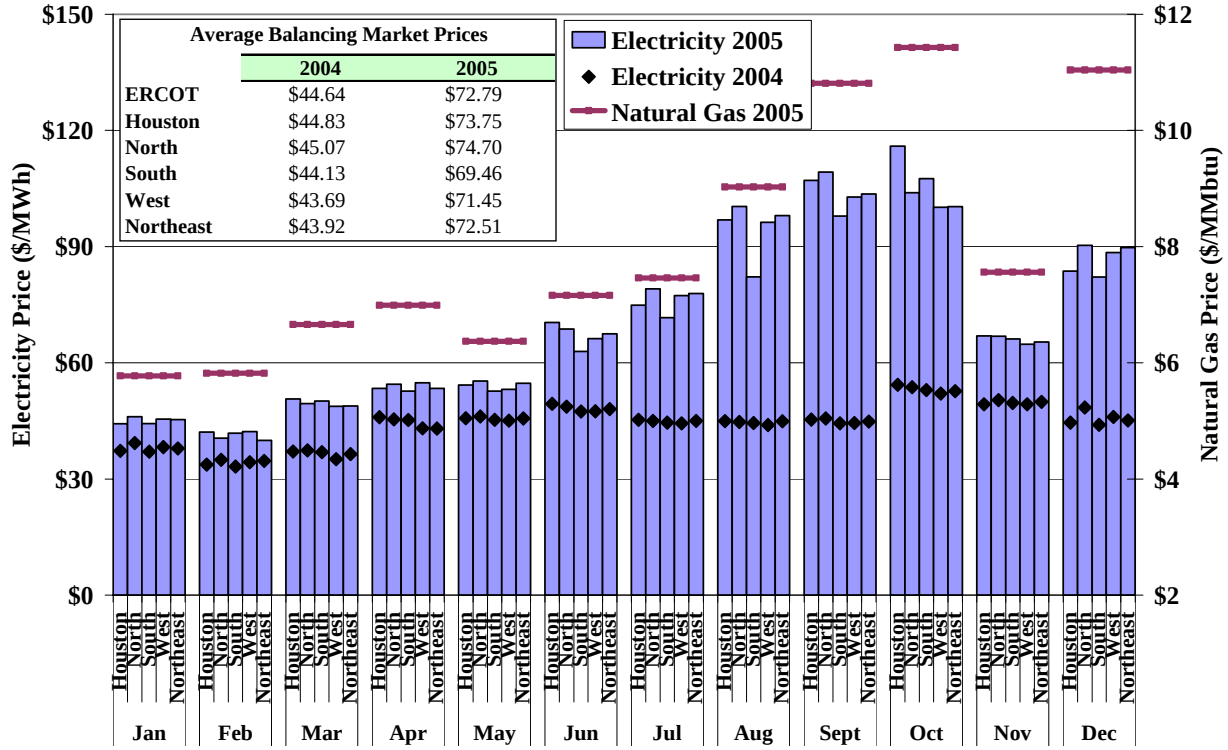
A. Review of Market Outcomes

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 the 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 2004 and 2005 and natural gas prices in 2005.

Balancing Energy Market Prices
2004 & 2005



Balancing energy prices were 63 percent higher on average in 2005 than in 2004, primarily due to a remarkable rise in natural gas prices. Prices began to rise significantly during the summer of 2005 and remained at high levels through the end of the year. These movements in energy prices were correlated with natural gas price movements. Natural gas prices rose rapidly in the late summer and peaked in September and December. The increase in natural gas prices was largely due to the effects of the hurricanes on the productive capability of the Gulf Cost region. In contrast to the results in 2005, balancing energy prices were relatively constant in 2004 due to lack of volatility in fuel prices.

Although fuel price fluctuations have been the dominant factor driving the increases in electricity prices in 2005, fuel prices alone do not explain all of the increases. At least two other factors contributed significantly to the higher prices during this period. First, ERCOT experienced substantially more super-peak demand hours (i.e. greater than 55 GW) than in previous years. ERCOT's current relatively high capacity margin limited the impact of higher demand on prices. Second, one supplier raised its offer prices significantly during the summer of 2005 relative to prior periods. These offer patterns are shown in Section V, which analyzes the competitive performance of the ERCOT market in 2005.

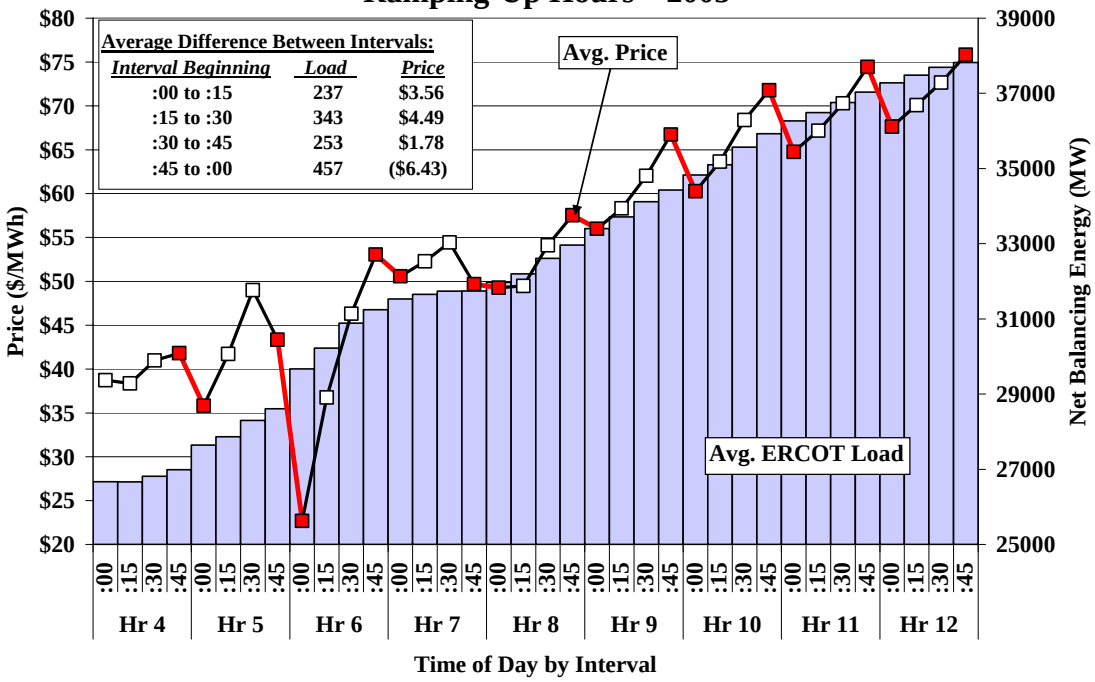
The figure above shows that transmission congestion between zones increased in ERCOT during 2005. The difference between the average North zone prices and the average South zone price was approximately 3.2 percent in 2004 and 7.5 percent in 2005. In individual months, the difference was much larger. For example, the average North zone price exceeded the South zone by approximately $18 per MWh in August 2005. The pattern of congestion has also changed. In 2004, congestion was most frequent on the South-to-North and Northeast-to-North Commercially Significant Constraints ("CSCs"). In 2005, the most substantial congestion was on the South-to-North, North-to-Houston, and South-to-Houston CSCs.

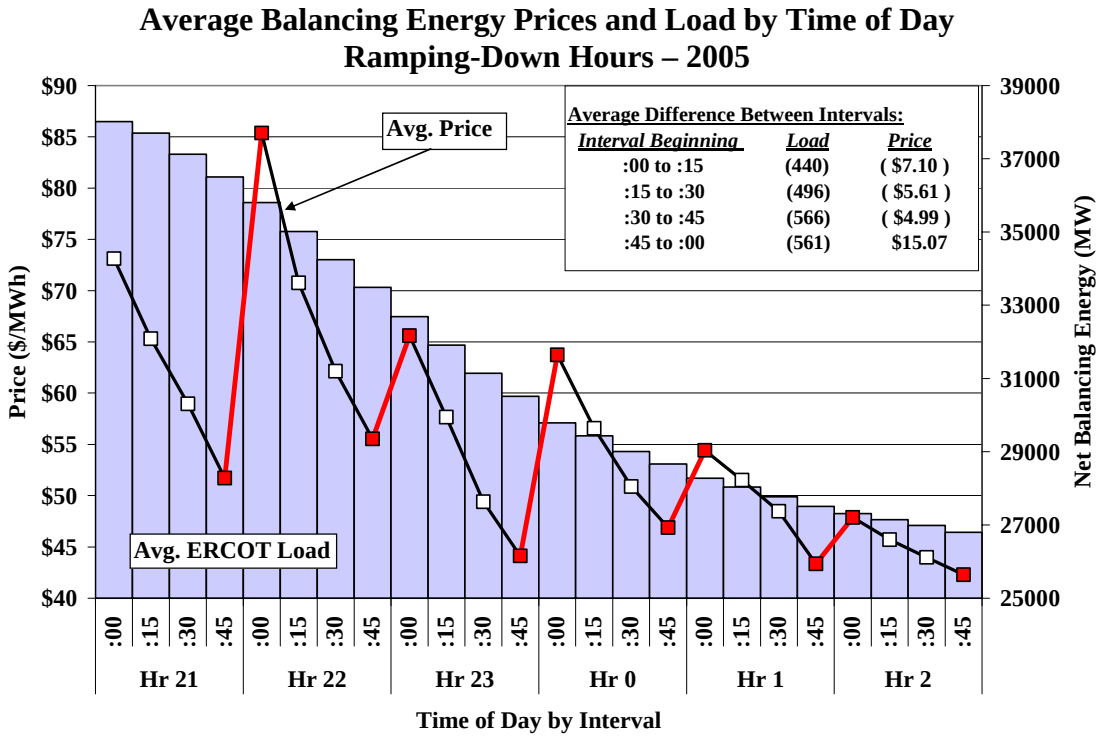
The report evaluates two other aspects of the balancing energy prices: 1) the correlation of the balancing energy prices with forward electricity prices in Texas, and 2) the primary determinants of balancing energy prices. Natural market forces should push forward market prices to levels consistent with expectations of spot market prices. Thus, it is surprising that average forward market prices were considerably lower than average balancing energy market prices in 2005.

Forward prices were comparable to balancing energy prices on the vast majority of days, although the forward market did not anticipate the extreme balancing prices that occurred on a small number of days but which had a significant impact on the average.

In addition to the factors discussed above, we have observed another significant issue related to the balancing energy prices. There is a clear relationship between the net balancing energy deployments and the balancing energy prices. This is not expected in a well-functioning market. The report concludes that this relationship is partly due to the hourly scheduling patterns of most of the market participants. 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 two 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 – 2005



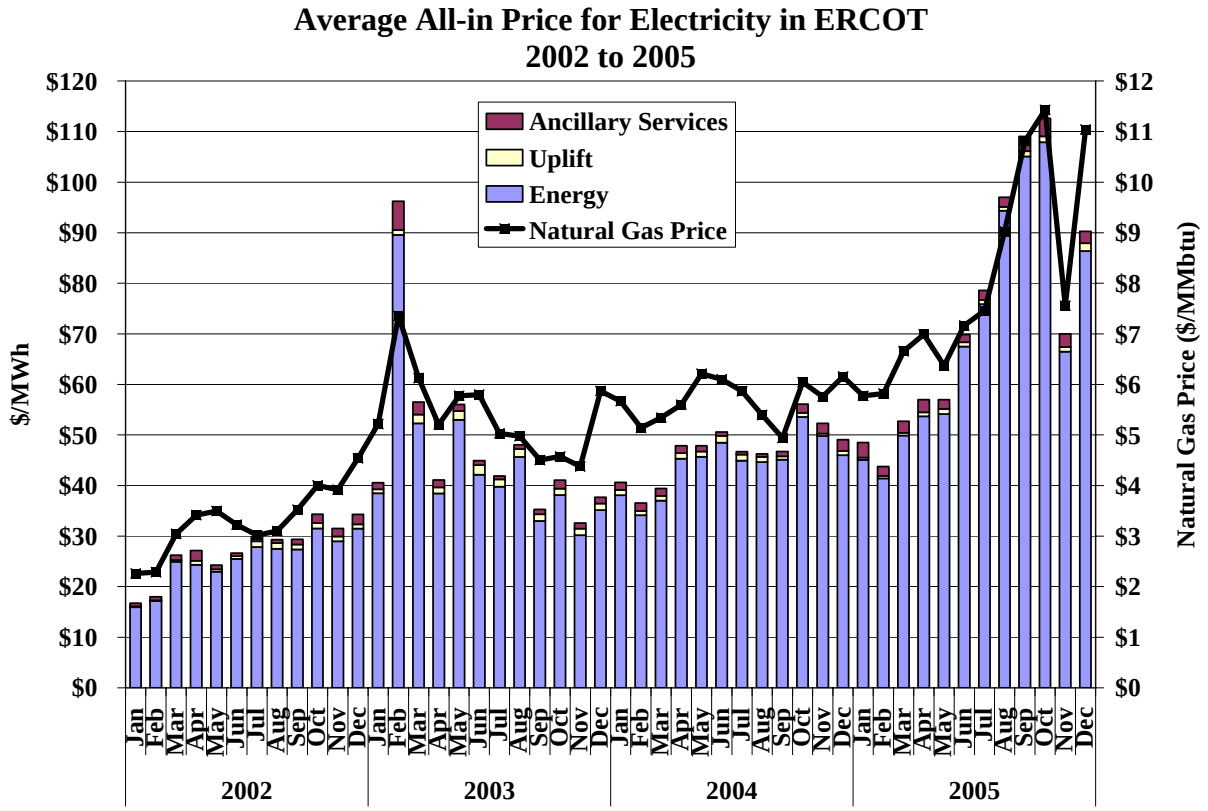


These pricing patterns raise significant efficiency concerns regarding the operation of the balancing energy market. Moreover, this pattern has been consistently observed for several years and is likely to continue until changes are made to the market rules.[2] This report includes several recommendations we have made to address the issue under the current zonal design. However, a more comprehensive solution to these operational issues is possible in the nodal framework which is scheduled for implementation in 2009.

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"). These costs, regardless of the location of the congestion, are borne equally by all loads within ERCOT. 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 four years are shown in the figure below along with a natural gas price trend.

[2] See 2003 SOM Report, Assessment of Operations, and 2004 SOM Report



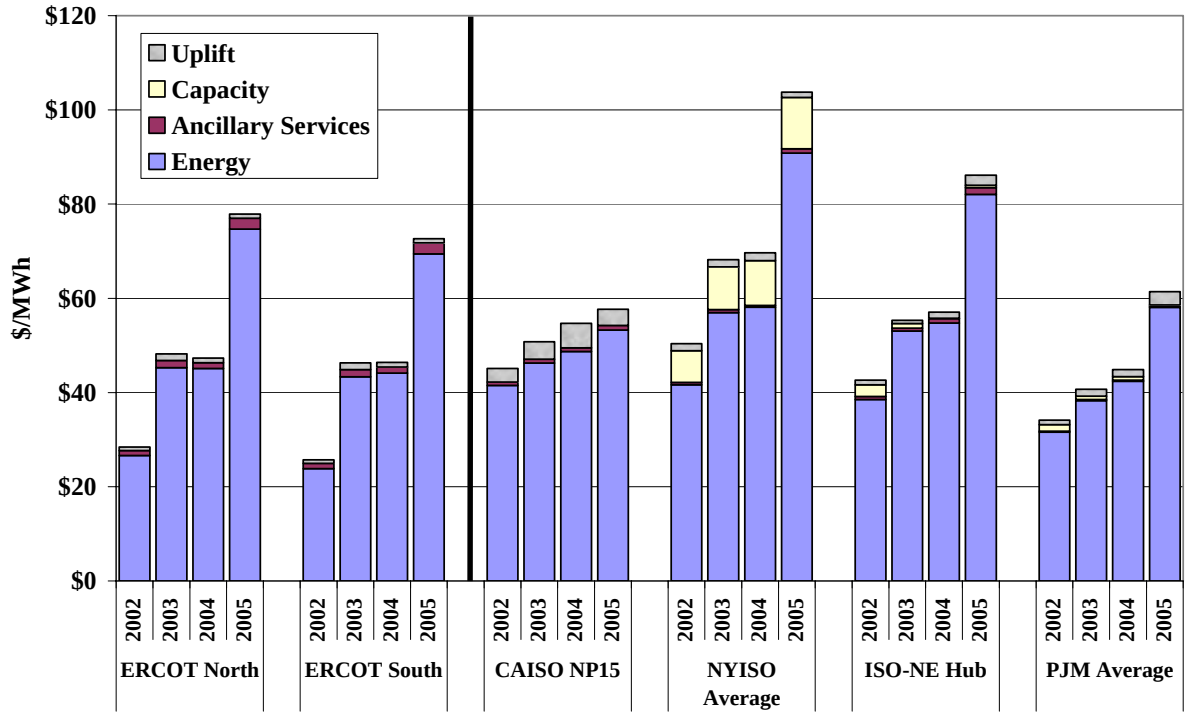
The figure indicates that natural gas prices were the primary driver of the trends in electricity prices from 2002 to 2005. Natural gas prices increased in 2003 by an average of more than 65 percent from 2002 levels while the all-in price for electricity increased by 72 percent in the same timeframe. In 2005, natural gas prices again increased sharply. Natural gas prices increased by an average of more than 41 percent from 2004 levels, contributing to an increase in the all-in price for electricity of 63 percent.

From 2004 to 2005, an 81 percent increase in ancillary services costs caused 2 percent of the increase in the all-in price for electricity. Ancillary services prices began to increase in the fall of 2004 and remained relatively high throughout 2005. The higher ancillary services prices coincided with more frequent price spikes in the balancing energy market, which is to be expected since the energy and ancillary services requirements are satisfied by the same resources. There was a slight reduction in total uplift costs for local congestion in 2005, which translated into a large reduction in the share of the all-in price related to uplift.

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. The following figure compares

the all-in prices for two ERCOT zones with four organized electricity markets in the U.S.: (a) California ISO, (b) New York ISO, (c) ISO New England, and (d) PJM. For each region, the figure reports the average cost (per MWh of load) for energy, ancillary services (reserves and regulation), capacity markets (if applicable), and uplift for economically out-of-merit resources.

Comparison of All-In Prices across Markets
2002 to 2005



Wholesale electricity markets in the U.S. experienced substantial increases in energy prices from 2002 to 2003 and from 2004 to 2005 due to increased fuel costs. Although the markets vary substantially in the portion of their generating capacity that is fueled by natural gas, these units are on the margin and setting the wholesale spot prices in a large share of the hours in each of the markets. The largest increases in electricity prices occurred in ERCOT, New York, and New England, indicating natural gas resources are on the margin more frequently in these markets than in PJM and California. Coal-fired generation is on the margin in a larger share of the hours in PJM, making prices in that market less sensitive to increases in natural gas prices. California's prices exhibit the weakest relationship to natural gas prices. A large share of California's electricity is produced by hydroelectric generation whose supply is heavily dependant on the quantity of rainfall each year.

3. Ancillary Services Markets

The primary market-based ancillary services in ERCOT are up regulation, down regulation, and responsive reserves. ERCOT may also procure non-spinning reserves as needed, which it did in approximately one quarter of the hours in 2005. QSEs may self-schedule ancillary services or purchase their required ancillary services through the ERCOT markets.

Ancillary services prices have risen considerably since 2002, consistent with long-term trends in natural gas and electricity prices. Because ancillary services 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. Providers of responsive reserves and regulation can incur opportunity costs when they reduce the output from economic units to make the capability available to provide these services. In late 2004, there was a large increase in ancillary services prices, coincident with an increase in the frequency of price spikes in the balancing energy market.

Although ancillary services prices have risen over the last few years, the impact has been partly mitigated by reductions in the required quantities of regulation. In 2002, ERCOT required approximately 3,000 MW of combined up and down regulation. By 2005, the requirement was reduced to an average of 1,950 MW during ramping hours and 1,400 MW during non-ramping hours. This has *directly* reduced regulation costs by reducing the overall quantity scheduled, either through bilateral arrangements or through the day-ahead auction. This has also *indirectly* reduced regulation costs by reducing the clearing prices of regulation that would have prevailed under higher demand levels for regulation.

In this report, we compare the amounts of capacity scheduled to provide operating reserves to the quantities of capacity that are actually available in real time. In generally, we find that the capacity available to provide reserves in real time far exceeds the quantities scheduled to meet the operating reserves requirements. Although we find that individual QSEs sometimes do not set aside sufficient capacity in real-time to satisfy their operating reserves obligations, the magnitude of these shortages is very small relative to the market's required quantities of operating reserves. Stakeholders are currently in the process of determining how to implement PRR 590, which directs ERCOT to monitor the compliance of individual QSEs with their

operating reserves obligations. We also note that QSEs are sometimes short of their responsive reserves obligations due to the provision that does not allow more than 20 percent of the capacity of a particular resource to provide responsive reserves.

In October 2005, ERCOT modified the day-ahead procurement process for ancillary services so the markets for regulation, responsive reserves, and non-spinning reserves clear simultaneously. By running a simultaneous auction for the four services, it is possible to clear the market with the least expensive set of available offers. This change is likely to result in more efficient prices for ancillary services since they will better reflect the opportunity costs of not providing the other services.

ERCOT continued to incur relatively high costs for reserves and regulation compared with other markets in 2005. This occurs even though sufficient excess capacity is usually on-line and available to the balancing market. 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 2005, 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 it results in an allocation of resources to provide ancillary services that is suboptimal. The only market with comparable responsive reserves prices is PJM, which also does not jointly-optimize the procurement of reserves and energy.

The current Nodal Protocols specify that energy and ancillary services will be jointly optimized in a centralized day-ahead market. This is likely to improve the overall efficiency of the day-ahead unit commitment. However, we also recommend the development of real-time markets that co-optimize ancillary services and energy. This recommendation differs from the sequential markets for real-time energy and ancillary services that are being proposed for the initial implementation of the nodal market. Co-optimized markets are superior to sequential markets because they:

- More efficiently schedule resources to provide ancillary services and energy. No model running in a day-ahead or hour-ahead timeframe can accurately predict conditions in real-time. Hence, such markets necessarily allocate resources to competing services inefficiently.

- Facilitate efficient pricing during shortage conditions (i.e., "scarcity pricing"). One of the most significant benefits of co-optimization is that it results in prices that reflect the trade-off between energy and reserves during shortage conditions.
- Reduce uplift payments to generators. If pricing is not efficient under shortage conditions and expensive generation is dispatched to meet the needs of the system, these expensive generators will need to be compensated through make-whole payments on a frequent basis.

4. Net Revenue Analysis

A final analysis of the outcomes in the ERCOT markets in 2005 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 to 2005 for four types of units, a natural gas combined-cycle generator, a simple-cycle gas turbine, a coal-fired steam turbine with scrubbers, and a nuclear unit. The net revenue increased significantly from 2002 to 2005, largely due to rising natural gas prices and more frequent price spikes in the balancing energy market. Net revenue has increased most significantly for the coal-fired and nuclear units. This is explained by the fact that these units have enjoyed relatively constant

production costs over the last four years while the production costs for the natural gas-fired resources have increased and partially offset the increase in energy prices.

To evaluate the net revenue levels, we compare the net revenue to the annualized generic costs of new generation. We recognize that specific projects have higher or lower costs than these generic estimates. For example, projects to upgrade existing facilities or to re-power existing generation may cost substantially less than building a new unit. This comparison of net revenue and investment costs shows that:

- Since 2003, natural gas prices have been in a range that could potentially make new coal and nuclear investment economically viable. However, this result must be tempered by the fact that there are likely costs associated with resources that are in excess of the generic cost estimates we used, such as the cost of nuclear waste disposal.

- In 2005, balancing energy prices rose more in percentage terms than natural gas prices, causing the net revenues for combined cycle resources and simple cycle gas turbines to be close to the levels that support new investment.

This second finding raises some concern because one would not expect the market to provide economic signals that would support new investment in natural gas generation when a surplus of generation is prevailing in the region. This result can be explained in large part by the substantial quantities of balancing energy that were available, but not offered in the balancing energy market, which led to frequent shortages of balancing energy and associated price spikes.

Under the nodal market design, it will be important to ensure that the market sends efficient signals for new investment. This is primarily accomplished in one of two ways:

- A capacity market; and/or
- Shortage pricing provisions to ensure that prices rise appropriately in the energy and ancillary services markets to reflect the true costs of shortages when resources are insufficient to satisfy both the energy and ancillary services requirements.

A market with one or both of these provisions can maintain adequate levels of supply without allowing market participants to artificially raise prices by withholding resources. Relying on large participants to raise prices by withholding in order to generate efficient long-term economic signals is inferior to relying on shortage pricing provisions because it is unlikely to provide efficient price signals in each location. One of the most significant benefits of co-optimizing energy with ancillary services in real-time is that it allows for efficient shortage

pricing. Clearing prices in a co-optimized market reflect the trade-offs between energy and ancillary services when they system is in shortage and the requirements of both markets. Such a market allows prices to rise sharply during periods of true shortage, providing the long-term economic signals that govern new investment and retirement decisions in generation and transmission facilities.

B. Demand and Resource Adequacy

1. Installed Capacity and Peak Demand

Since electricity cannot be stored, the electricity market must ensure that generation matches load on a continuous basis. Thus, one critical issue for a wholesale electricity market is whether sufficient supplies exist to satisfy demand under peak conditions. In 2005, the load served by ERCOT reached a peak of nearly 61 GW.[3] This was not a particularly large increase relative to previous years since the peak was approximately 60 GW in 2003 and 59 GW in 2004. 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 2005 due to ERCOT's substantial 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.

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 86 percent in the Houston Zone. This makes ERCOT particularly vulnerable to natural gas price increases because the other resource types (coal and nuclear) are primarily base load units that are generally not the marginal source of supply.

When import capability, resources that can be switched to the SPP, and Loads acting as Resources are included, ERCOT has more than 80 GW of installed capacity. However, significant amounts of this are not kept constantly in service. ERCOT estimates that more than 5 GW was mothballed during 2005 and a large amount of capacity is used to satisfy cogeneration demands rather than to produce electricity. Furthermore, ambient temperature restrictions

[3] Includes transmission losses

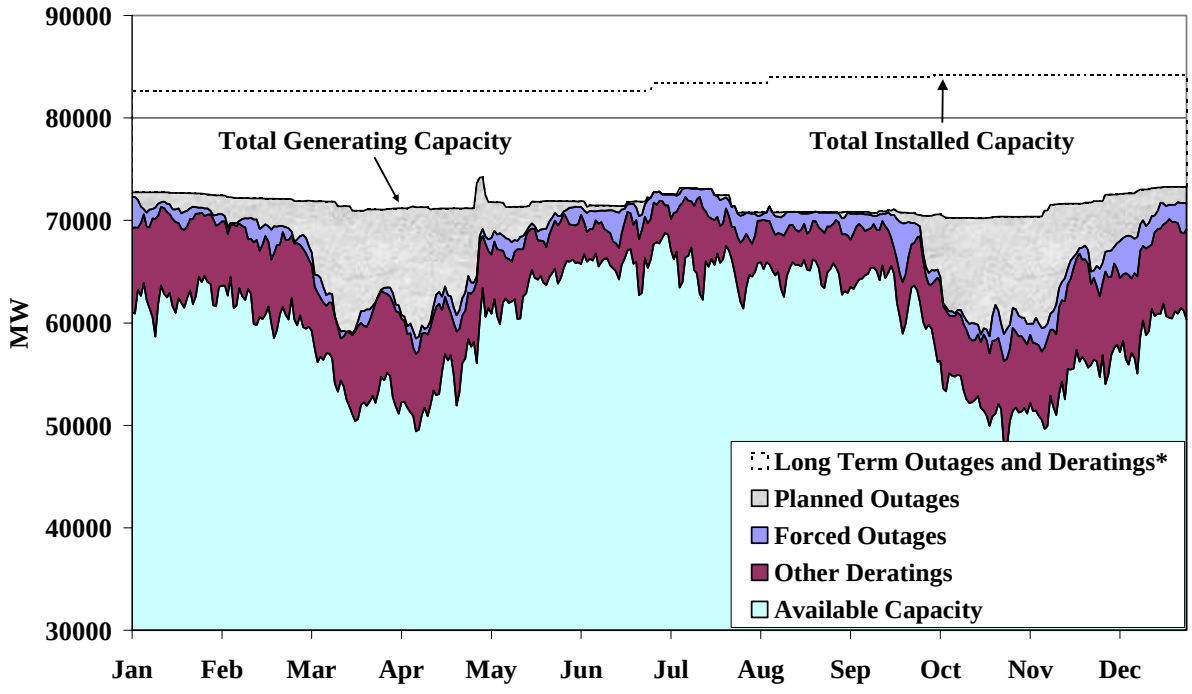
increase during the summer months when demand is highest, leading to substantial deratings. Although ERCOT had sufficient capacity to meet load and ancillary services needs during the 2005 peak, it is important to consider that electricity demand will continue to grow and that a significant number of generating units in Texas are soon reaching or are already exceeding their expected lifetimes. Without significant capacity additions, these factors may cause the resource margins in ERCOT to diminish rapidly over the next three to five years.

2. Generator Outages and Commitments

Despite adequate installed capacity, 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 the installed capability of a generating resource and its maximum capability (or "rating") in a given hour. Generators can be fully derated (rating equals

Short and Long-Term Deratings of Installed Capability
2005



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

** Switchable capacity is included under installed capacity in this figure.

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 previous figure shows the daily available and derated capability of generation in ERCOT.

The figure shows that long-term outages and deratings typically range from 10 GW to 14 GW. 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);
- Cogeneration resources typically used for purposes other than electricity generation; 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 2005 to as high as 11 percent in other months. These deratings include outages not reported or correctly logged by ERCOT and natural deratings due to high ambient temperature conditions and other factors. The overall pattern of outages and deratings is consistent with competitive expectations and does not raise any significant concerns.

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, responsive reserves provided by generation, 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 more than 4 GW in 2005, although substantially reduced from 2004 when it averaged nearly 7 GW. These results show that the ERCOT system is still generally over-committed, indicating 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 a day-ahead energy market, which is planned under the current nodal market design, promises substantial efficiency improvements in the commitment of generating resources.

3. 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.

During 2005, 92 resources totaling 1,835 MW of capability were qualified as LaaRs. The amount of responsive reserves provided by LaaRs gradually increased from about 900 MW at the beginning of 2004 to an average of 1,133 MW in 2005. Currently, LaaRs are permitted to supply up to 1,150 MW, 50 percent, of the responsive reserves requirement. Although the participants with LaaRs resources are qualified to provide non-spinning reserves and balancing up load in real-time, in 2005, they provided less than 2 percent of non-spinning reserves and none of the balancing energy. This is not surprising because the value of curtailed load tends to be relatively high, and providing responsive reserves offers substantial revenue with very little probability of being deployed. In contrast, resources providing non-spinning reserves are 70 times more likely to be deployed. In addition, 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 non-spinning reserves or balancing energy.

The clearing price for responsive reserves provided by LaaRs is set by the marginal generator, although the quantity of LaaRs willing to supply responsive reserves at the clearing price

exceeds the demand (i.e. 1,150 MW). This design encourages inefficient behavior by QSEs that want to sell responsive reserves from their demand resources. In order to be selected, LaaRs must submit an offer price that is among the lowest priced 1,150 MW of LaaRs. This gives QSEs an incentive to offer LaaRs at very low (and even negative) prices. To improve the efficiency of responsive reserves prices and incentives for suppliers, we recommend that ERCOT consider setting separate prices for different types of responsive reserves whenever the 1,150 MW limitation on LaaRs prevents more from being scheduled. This would improve incentives for market participants, lead to a more efficient selection of resources, and reduce ancillary services costs for consumers. This is consistent with a recommendation made by a Special Task Force of the Demand Side Working Group to deal with this issue that was endorsed by the Wholesale Market Subcommittee.[4]

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.

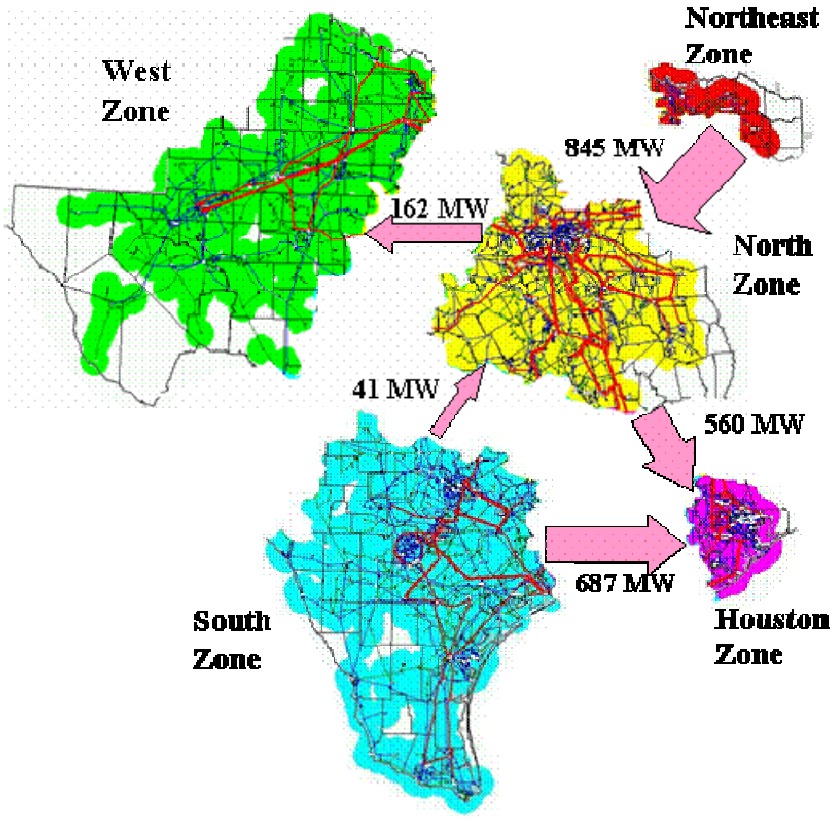
1. Electricity Flows between Zones and Interzonal Congestion

The balancing energy market uses the Scheduling, Pricing, and Dispatch ("SPD") software which 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 five zone-based locations and six transmission interfaces. The transmission interfaces are referred to

[4] See minutes of the March 22, 2006 Wholesale Market SubCommittee meeting

as Commercially Significant Constraints ("CSCs"). The following figure shows the average flows modeled in SPD during 2005 over each of these CSCs.

Average Modeled Flows on Commercially Significant Constraints
2005



Note: In the figure above, CSC flows are averaged taking the direction into account. So one arrow shows the average flow for the North-to-West CSC was 162 MW, while the average for the West-to-North CSC was *negative* 162 MW.

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

- 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 a CSC. When these flows cause congestion, it is beneficial to create a new CSC, such as the North to West CSC that was implemented by ERCOT in 2005 to better manage congestion over that path.
- Based on modeled flows, Houston is a significant importer while the Northeast Zone and the South Zone export significant amounts of power.

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 rose considerably to $119 million in 2005, which reflects an increase of $78 million from 2004. This increase was the result of more frequent congestion on the South-to-Houston, North-to-Houston, and South-to-North CSCs, as well as higher overall prices.

To account for the fact that the modeled flows can vary substantially from the actual physical flows (due to the simplifying assumptions in the model), ERCOT operators must adjust the modeled limits for the CSC interfaces to ensure that the physical flows do not exceed the physical limits. This process results in highly variable limits in the market model for the CSC interfaces.

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. Because the modeled limits for the CSC interfaces vary substantially, the quantity of TCRs defined over a congested CSC frequently exceeds the modeled limits for 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. The aggregate shortfall rose considerably to $38 million in 2005, up from $10 million in 2003 and $8 million in 2004.

The pricing of the congestion rights is also important because the revenue from the auction of the congestion rights is the 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 2005, the annual and monthly TCR auctions substantially under-valued the TCRs in comparison to the balancing market congestion. While payments to TCR holders totaled $119 million, the

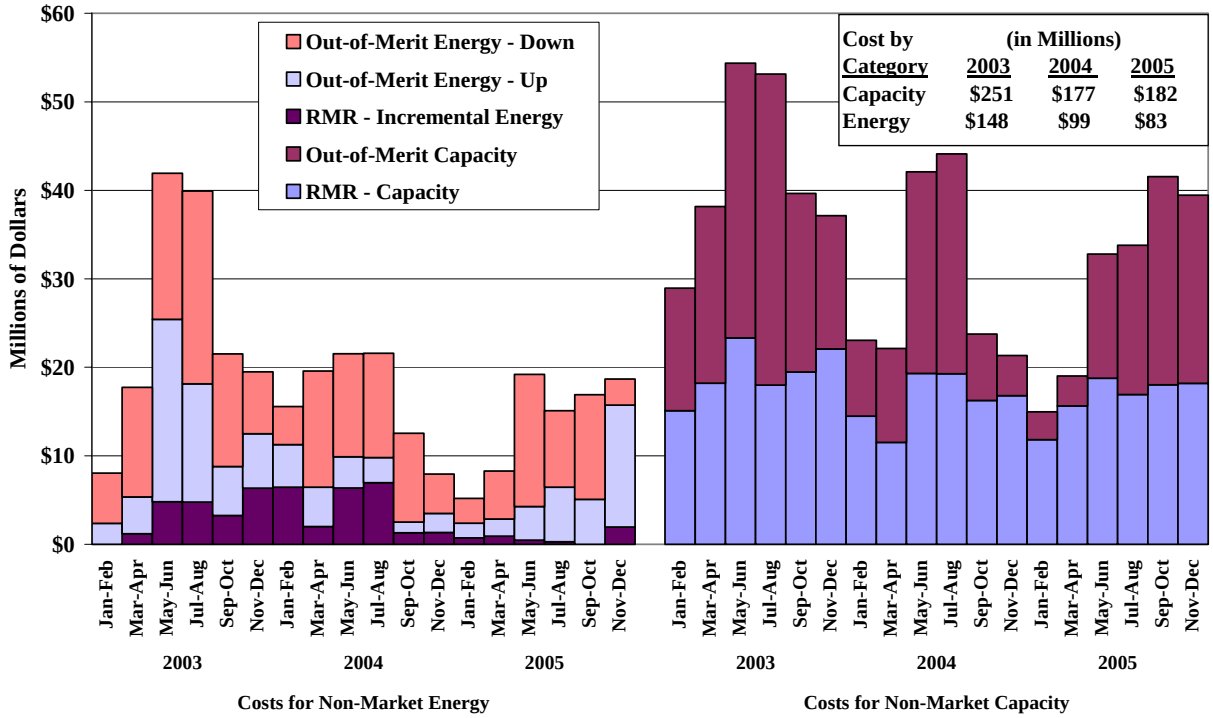
holders of these rights paid approximately $40 million for them in the TCR auctions, which was more consistent with the payments to TCR holders in the two previous years.

2. 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 relevant local area. 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, from 2003 to 2005.

The figure shows that OOME costs and incremental energy costs from RMR units declined from $148 million to $83 million from 2003 to 2005, a decrease of 44 percent. Likewise, the costs of OOMC commitments and the capacity costs from RMR units declined 27 percent from 2003 to 2005. The decline in uplift was primarily attributable to sizable reductions in payments for OOME-Up, OOME-Down, and OOMC commitments. The figure also shows that all classes of out-of-merit costs tend to increase 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, the seasonal pattern changed in the fall of 2005 as uplift costs remained elevated after the summer. The higher uplift was due to larger quantities of OOMC and OOME in the South and West zones. It is likely that expenses for local congestion management would have decreased even more if natural gas prices had not risen dramatically in the fall of 2005.

Expenses for Out-of-Merit Capacity and Energy
2003-2005



The report finds four primary factors that contributed to the reduction of these out-of-merit costs. First, the addition of the Northeast Zone at the beginning of 2004 allowed a significant amount of congestion that had been local to become interzonal congestion between the Northeast and revised North zones in 2004. This change reduced the OOME Down dispatch within the Northeast and the OOME Up in Dallas-Fort Worth ("DFW") and other areas within the North zone. Second, the definition of an additional CSC from the North zone to Houston reduced local congestion by allowing the zonal energy market to manage the congestion on the transmission facilities connecting the two areas. Third, the formula for OOMC payments was revised, which reduced the incentive for suppliers to wait for ERCOT to commit their units through the OOMC process on days when the units would otherwise be economic. Fourth, in March 2005, ERCOT began using dynamic ratings for many transmission facilities that are frequently congested. By using real-time information for each facility, ERCOT is able to more fully utilize the transfer capability of the transmission system while maintaining reliability.

3. Conclusions regarding Transmission Congestion in ERCOT

The results in this area of the report indicate that:

- ERCOT has taken steps that have helped reduce annual intrazonal congestion costs by $134 million since 2003;
- Although inter-zonal congestion rose considerably in 2005, the vast majority of congestion in ERCOT is still intrazonal. The uplift associated with intrazonal congestion is difficult for loads to hedge and is not transparent;
- The current zonal market can result in large inconsistencies between the interzonal flows calculated by SPD and the actual flows over the CSC interfaces, leading to $38 million in uplift costs in 2005; and
- These inconsistencies can result in under-utilized transmission capability and difficulties in defining transmission rights whose obligations can be fully satisfied.

The long-run remedy for both the interzonal and intrazonal issues identified in this report will be the implementation of nodal markets. The nodal markets currently being designed for implementation in 2009 will provide transparent prices for both generators and loads that would fully reflect all transmission constraints on the ERCOT network.

D. Balancing Energy Offers and Schedules

QSEs play an important role in the current ERCOT markets. QSEs must submit balanced schedules so that the quantity of generation scheduled matches the quantity of load scheduled prior to real-time. However, there is no longer a requirement for the scheduled load to match the forecast of real-time load. When actual real-time load exceeds the energy scheduled prior to real-time, the remaining load is served by energy purchased in the balancing energy market. Conversely, when scheduled energy exceeds actual real-time load, load serving entities sell their excess to the balancing energy market. QSEs submit balancing energy offers to increase or decrease their energy output from the scheduled energy 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. Hourly Schedule 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:

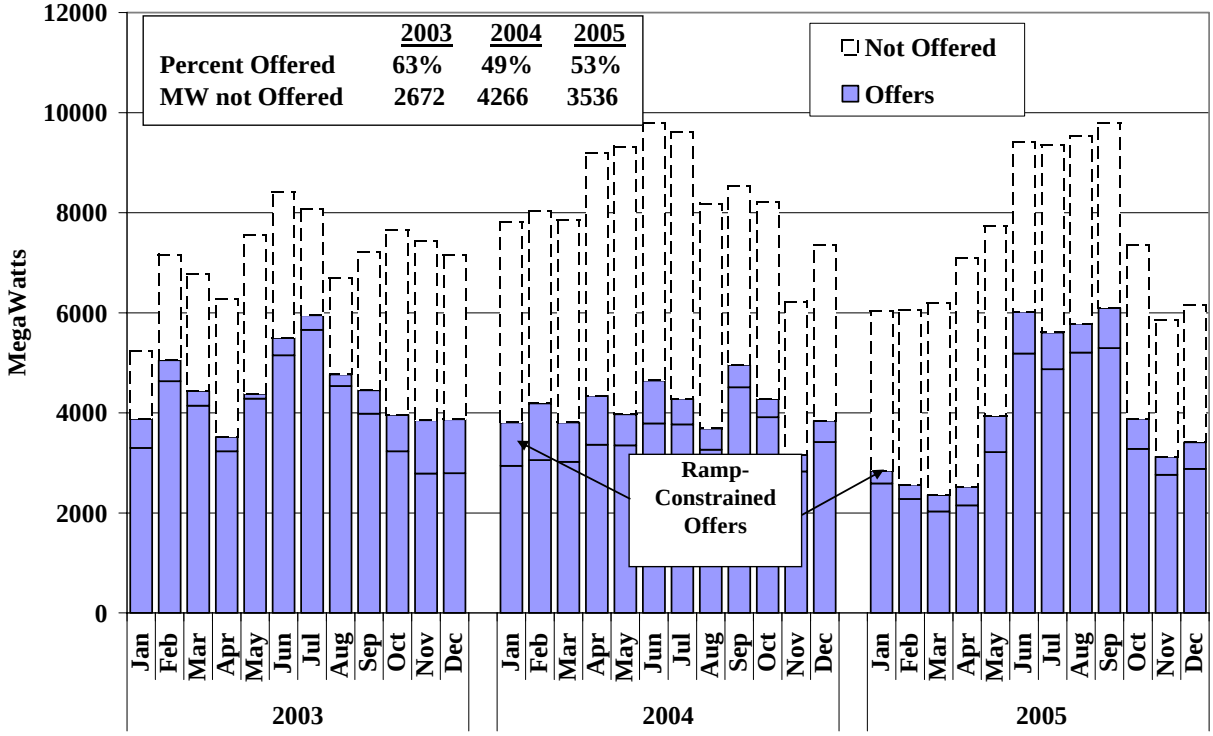
- In these ramping hours, the loads are generally moving approximately 300 to 500 MW each 15-minute interval.
- Although QSE's can modify their schedules each interval, most only change their schedules hourly, resulting in schedule changes averaging 1000 to 4000 MW in these hours (and sometimes significantly 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 most of 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). These recommendations were considered by ERCOT and its participants, but were not proposed in any subsequent PRR. However, the issues that these recommendations were designed to address should be resolved by the implementation of unit-specific dispatch under the nodal market design.

2. 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 on average in each month from 2003 to 2005.

Available Balancing Energy vs. Balancing Energy Offers
Daily Peak Load Hours – 2003 to 2005



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

- In general, approximately 53 percent of the available energy is offered in the balancing energy market.
- The largest QSEs offer a larger share of their available energy than smaller suppliers.
- 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).

While it is a significant concern that not all available capacity is offered into the market, the results of our analyses do not suggest that the large amount of un-offered capacity represents strategic withholding.[5] Rather, the report identifies several issues that might prevent market participants from offering all of their available capability from online or quick-starting resources. In particular, the current market rules and portfolio bidding framework results in ramp limitations that are much tighter than the true physical ramp limitations of the individual generating units. This reduces the ability of the market to fully utilize the generating resources and can result in inefficient transitory fluctuations in balancing energy prices.

[5] This does not rule out the possibility that strategic withholding occurred through some other means.

The report includes a number of recommendations to address the portfolio ramp limitations and better allow gas turbine capacity and duct firing capacity to be included in the portfolio offers. These recommendations are:

- Consider the feasibility of allowing QSEs to offer multiple ramp rates that vary by output level. A Protocol Revision Request has been drafted to accomplish this under the current market design, and it is being considered by the Wholesale Market Subcommittee.
- Modify the treatment of ramp limitations in the balancing energy market to recognize ramping capability that is used/made available associated with QSEs' schedule changes. This will be addressed when ERCOT implements the nodal market design with unit-specific dispatch.

3. Resource Plan Analysis

QSEs submit resource plans to inform ERCOT about which resources they plan to use to satisfy their energy and ancillary services obligations. While QSEs are expected to make their best effort to accurately forecast how they will operate their units, the resource plans are not financially binding. Resource plans are used by ERCOT in some of its reliability assessments before real-time and to make additional commitments to maintain reliability. Therefore, it is important for ERCOT to have accurate information in the resource plans that QSEs submit in order to avoid taking unnecessary and sometimes costly actions to maintain reliability.

Resource plans are not financially binding, yet they are used by ERCOT to make commitment decisions that can have significant cost implications. Hence, 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. We analyzed market participants' resource plans to evaluate whether the market protocols may provide incentives for such strategic conduct. Specifically, we evaluated units that are frequently committed out-of-merit or frequently dispatched out-of-merit. Such units receive additional payments from ERCOT and we investigated whether market participants may engage in strategies to increase these payments. This analysis provides evidence that market participants have engaged in strategies that increase:

- *OOMC Commitment* – Our analysis 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 behavior forces ERCOT to make more OOMC commitments, resulting in higher local congestion uplift costs.
- *OOME Up Dispatch* – QSEs with resources that frequently receive OOME Up instructions typically under-schedule these resources in the final real-time resource plan.

This behavior leads ERCOT to deploy these resources upward for local congestion management, resulting in higher local congestion uplift costs.

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. This stems from the lack of nodal prices to signal the value of capacity and energy in local areas. In the absence of nodal prices, market participants may act strategically to garner additional uplift 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.

The pivotal supplier analysis indicates that the two largest suppliers in ERCOT were frequently pivotal in 2005. Moreover, because a large portion of the available energy from online resources was routinely not offered in the balancing energy market, we found that a supplier was pivotal more frequently if we only consider resources that were offered to the balancing energy market. Although additional supply was capable of entering the balancing market, several factors described above may prevent full utilization of the available energy in ERCOT making the balancing market more vulnerable to manipulation. In this analysis, we found that the frequency of a supplier being pivotal increased with the level of demand.

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 analyzed measures of physical and economic withholding in order to further evaluate competitive performance of the ERCOT market. Withholding patterns were examined relative to the level of demand and the size of each supplier's portfolio. Based on the analyses conducted in this area, the report found patterns suggestive of economic withholding that raise competitive concerns.

One company had a larger share of unutilized economic capacity than the other suppliers in ERCOT. This unutilized capacity grew at higher load levels when the effect on prices was

generally highest. Although we did not analyze the impact on balancing market prices for this report, the quantity of unutilized capacity may be large enough to have had a significant impact on prices during the summer months.

I. REVIEW OF MARKET OUTCOMES

A. Balancing Energy Market

1. Balancing Energy Prices During 2005

The balancing energy market is the spot market for electricity in ERCOT. As is typical in other wholesale markets, only a small share of the power produced in ERCOT is transacted in the 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 (including bilateral 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).[6] Hence, artificially-low prices in the balancing energy market will translate to artificially-low forward prices. Likewise, price spikes in the balancing energy market will increase prices in the forward markets. The analyses in this section summarize and evaluate the prices that prevailed in the balancing energy market during 2005.

Balancing energy market prices were considerably higher in 2005 than in 2004, particularly during the latter half of the year. The differences were primarily due to fluctuations in natural gas prices, which were unusually volatile from July 2005 through the end of the year. To summarize the price levels during the past two years, Figure 1 shows the load-weighted average balancing energy market prices in each of the ERCOT zones in 2004 and 2005.[7]

The table in Figure 1 indicates that balancing energy prices were 63 percent higher for ERCOT in 2005 than in 2004. Prices began to rise significantly during the summer of 2005 and remained

---

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

7 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 average prices reported elsewhere that are weighted by the balancing energy procured in the interval, which is a methodology we use to evaluate certain aspects of the balancing energy market. For this evaluation, balancing energy prices are load-weighted since this is the most representative of what loads are likely to pay (assuming that balancing energy prices are generally consistent with bilateral contract prices).

at high levels through the end of the year. These movements in energy prices were correlated with natural gas price movements. Natural gas prices rose rapidly in the late summer and peaked in September and December. The increase in natural gas prices was largely due to the effects of the hurricanes on the productive capability of the Gulf Cost region. In contrast to the results in 2005, balancing energy prices were relatively constant in 2004 due to lack of volatility in fuel prices.

Figure 1: Average Balancing Energy Market Prices
2004 & 2005

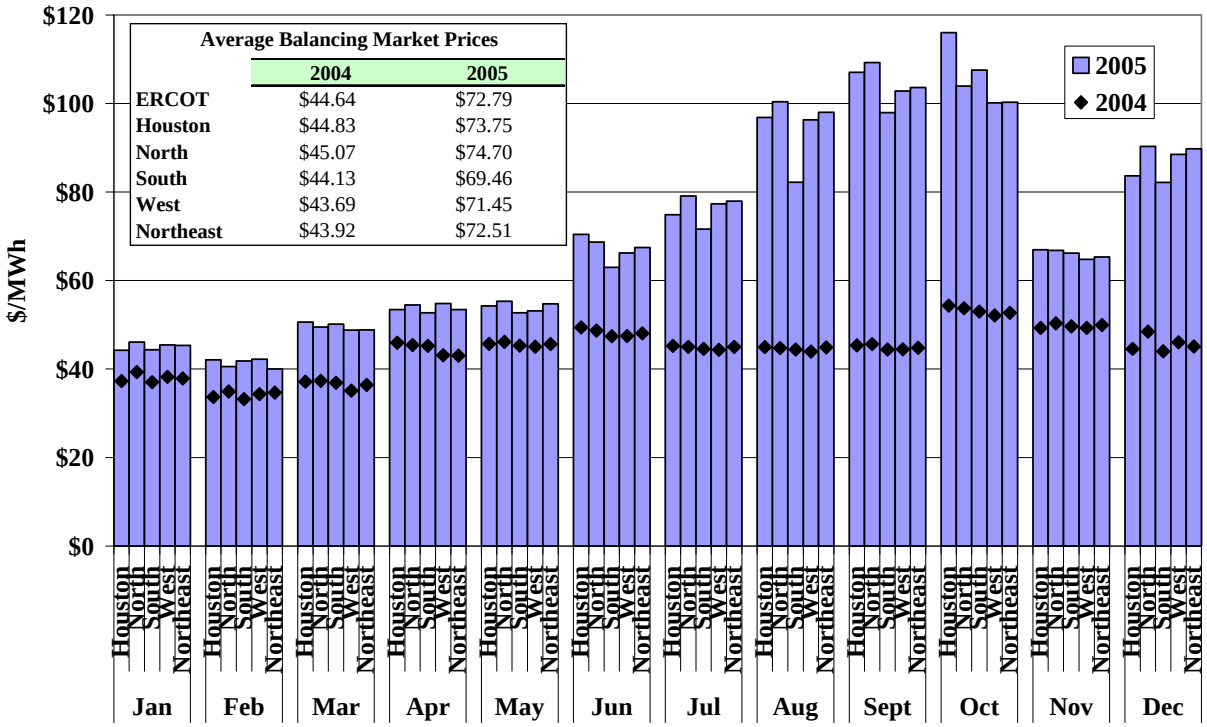


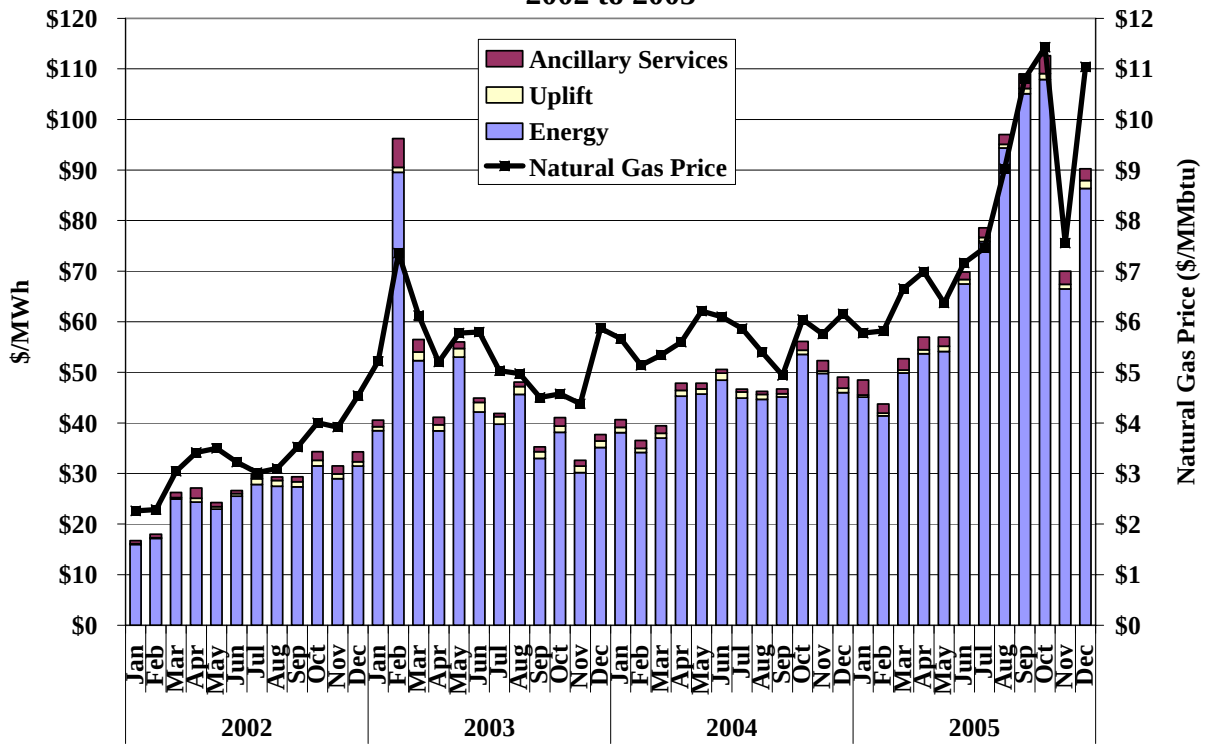
Figure 1 also shows that transmission congestion between zones increased in ERCOT during 2005. The difference between the average North zone prices and the average South zone price was approximately 3.2 percent in 2004 and 7.5 percent in 2005. In individual months, the difference was much larger. For example, the average North zone price exceeded the South zone by approximately $18 per MWh in August 2005. The pattern of congestion has also changed. In 2004, congestion was most frequent on the South-to-North and Northeast-to-North Commercially Significant Constraints ("CSCs"). In 2005, the most substantial congestion was on the South-to-North, North-to-Houston, and South-to-Houston CSCs.

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".[8] 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. Figure 2 shows the monthly average all-in price for all of ERCOT from 2002 to 2005.

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

- Energy costs: Balancing energy market prices are used to estimate energy costs, under the assumption that the price of bilateral energy purchases converges with balancing energy market 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 2005



[8] As discussed more below, uplift costs are costs that are allocated to load that pay for out-of-merit dispatch, out-of-merit commitment, and Reliability-Must-Run contracts.

Figure 2 indicates that natural gas prices were a primary driver of the trends in electricity prices from 2002 to 2005. 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 market prices. Natural gas prices increased in 2003 by more than 65 percent from 2002 levels on average while the all-in price for electricity increased by 72 percent. Again, natural gas prices increased in 2005 by an average of more than 41 percent from 2004 levels while the all-in price for electricity increased by 63 percent.

Although fuel price fluctuations have been the dominant factor driving the increases in electricity prices in 2005, fuel prices alone do not explain all of the increases. At least three other factors contributed significantly to the higher prices during this period. First, ERCOT experienced substantially more super-peak demand hours (i.e. greater than 55 GW) than in previous years. ERCOT's current relatively high capacity margin limited the impact of higher demand on prices. Second, less excess committed capacity occurred in 2005 on average.[9] Third, a supplier raised its offer prices significantly during the summer of 2005 relative to prior periods. Section III of this report discusses the first two factors in greater detail, while the third factor is examined in Section V. Analyses in the next sub-section adjust for natural gas price fluctuations to better highlight variations in electricity prices not related to fuel costs.

From 2004 to 2005, an 81 percent increase in ancillary services costs caused 2 percent of the increase in the all-in price for electricity. Ancillary services prices began to increase in the fall of 2004 and remained relatively high throughout 2005. The higher ancillary services prices coincided with more frequent price spikes in the balancing energy market, which is to be expected since the energy and ancillary services requirements are satisfied by the same resources. There was a slight reduction in total uplift costs for local congestion in 2005, which translated into a large reduction in the share of the all-in price related to uplift.

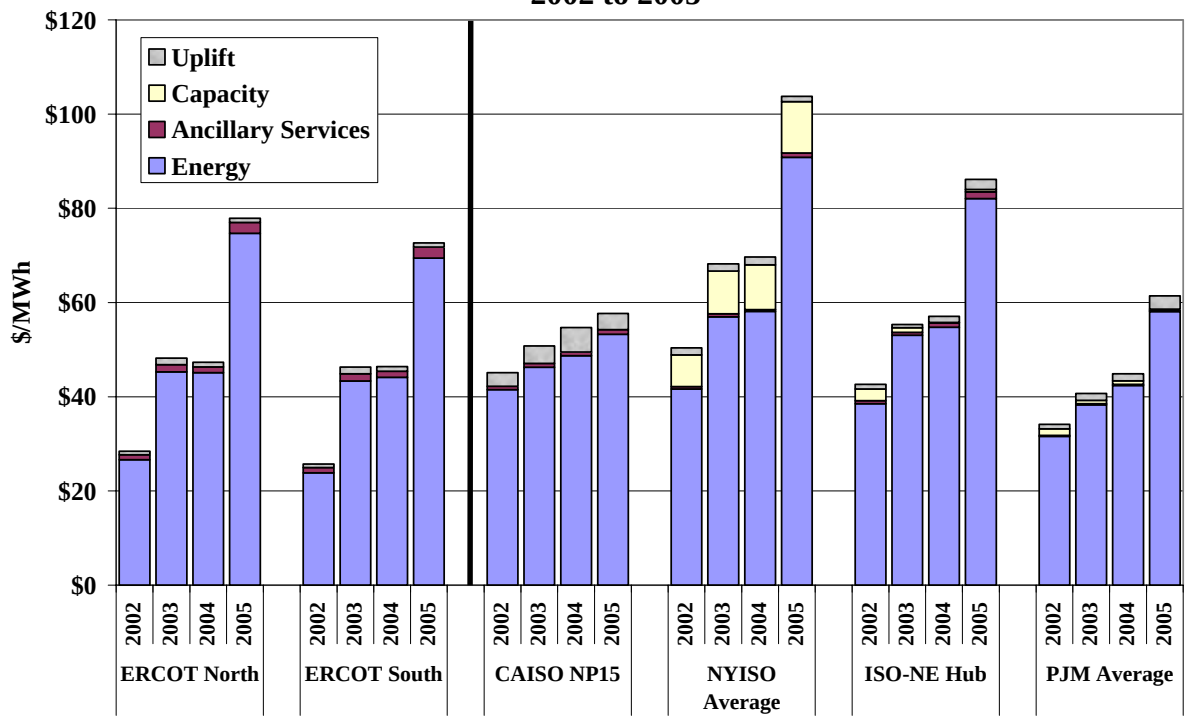
While all-in prices have generally moved in proportion with natural gas prices, a notable exception to the trend was in February 2003 when the rise in balancing energy prices far

[9] Excess committed capacity is total on-line generation minus peak real-time demand plus responsive reserves. It represents the amount of generation that could have remained off-line with ERCOT still having the ability to satisfy its energy and responsive reserve requirements.

exceeded the rise in fuel prices. 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.

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. The following figure compares the all-in prices for two ERCOT zones with four organized electricity markets in the U.S.: (a) California ISO, (b) New York ISO, (c) ISO New England, and (d) PJM. For each region, the figure reports the average cost (per MWh of load) for energy, ancillary services (reserves and regulation), capacity markets (if applicable), and uplift for economically out-of-merit resources.

Figure 3: Comparison of All-in Prices Across Markets 2002 to 2005

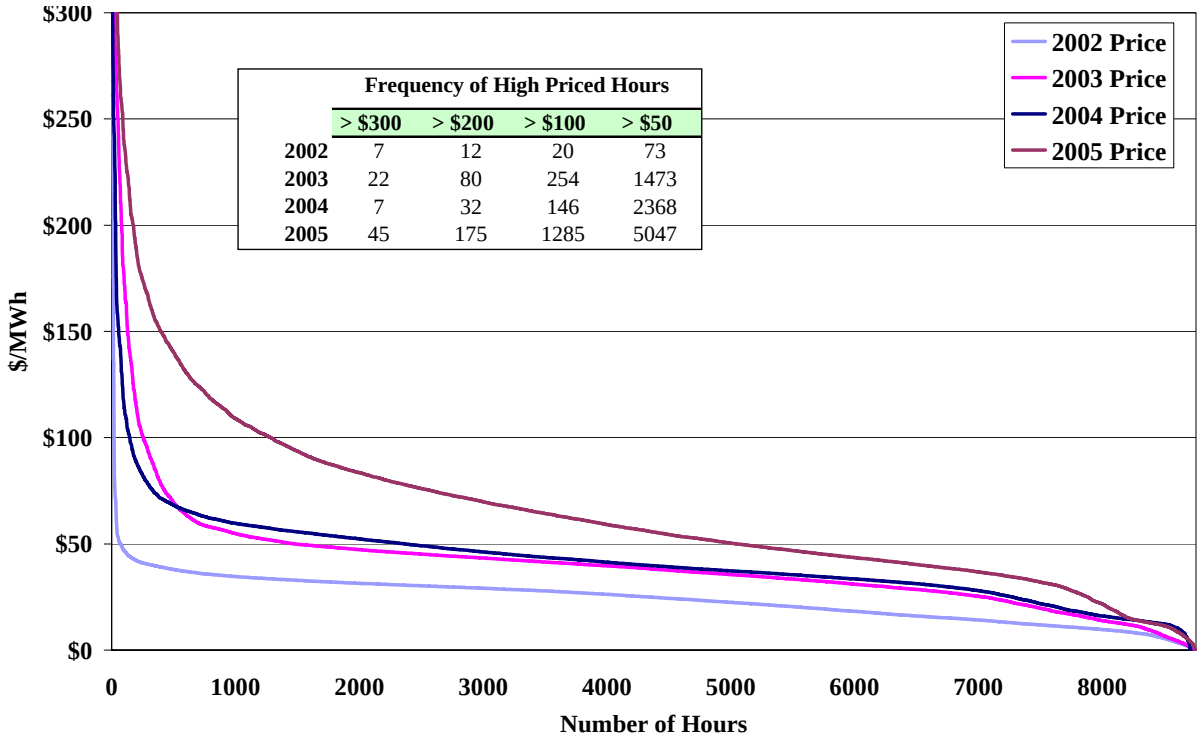


Wholesale electricity markets in the U.S. experienced substantial increases in energy prices from 2002 to 2003 and from 2004 to 2005 due to increased fuel costs. Although the markets vary substantially in the portion of their generating capacity that is fueled by natural gas, these units are on the margin and setting the wholesale spot prices in a large share of the hours in each of the markets. The largest increases in electricity prices occurred in ERCOT, New York, and New England, indicating natural gas resources are on the margin more frequently in these markets

than in PJM and California. Coal-fired generation is on the margin in a larger share of the hours in PJM, making prices in that market less sensitive to increases in natural gas prices. California's prices exhibit the weakest relationship to natural gas prices. A large share of California's electricity is produced by hydroelectric generation whose supply is heavily dependant on the quantity of rainfall each year.

Figure 4 presents price duration curves for the balancing energy market in each year from 2002 to 2005. A price duration curve indicates the number of hours (shown on the horizontal axis) that the price is at or above a certain level (shown on the vertical axis). The prices in this figure are hourly load-weighted average prices for the ERCOT balancing energy market.

Figure 4: ERCOT Price Duration Curve
2002 to 2005



Frequency of High Priced Hours				
	> $300	> $200	> $100	> $50
2002	7	12	20	73
2003	22	80	254	1473
2004	7	32	146	2368
2005	45	175	1285	5047

The figure shows that prices were considerably higher in 2005 than in previous years. For instance, balancing energy prices exceeded $50 in more than 5000 hours in 2005 and less than 2400 hours in 2004. While prices were comparable between 2003 and 2004, they were much higher than in 2002 when prices exceeded $50 in just 73 hours. These large year-to-year changes reflect the effects of higher fuel prices, which impact electricity prices in a broad range

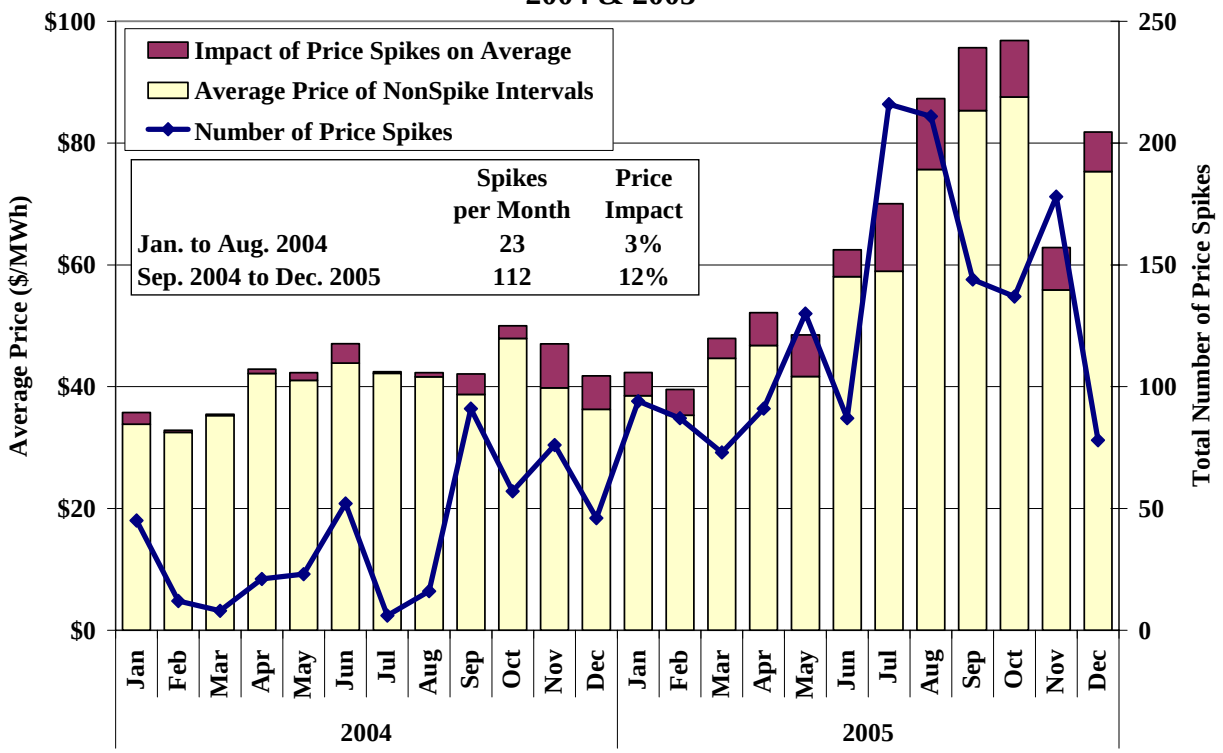
of hours. Higher natural gas prices raise the marginal production costs of the generating units that set the prices in the balancing energy market in a large share of the intervals.

Other market factors that affect balancing energy prices occur in a subset of intervals, such as the extreme demand conditions that occur during the summer. Figure 4 shows that there were differences in balancing energy market prices between 2002 and 2005 at the highest price levels. For instance, 2003 experienced considerably more price spikes (e.g., prices higher than $300) than 2004 even though prices were higher on average in 2004. 2005 exhibited the largest number of price spikes of any year due to the high demand levels in the summer and the substantial quantities of available balancing energy not offered in the balancing energy market.

To better observe the highest-priced hours during 2004 and 2005, the following analysis focuses on the frequency of price spikes in the balancing energy market. Figure 5 shows average prices and the number of price spikes in each month of 2004 and 2005. In this case, price spikes are defined as intervals where the load-weighted average Market Clearing Price of Energy ("MCPE") in ERCOT is greater than 18 MMbtu per MWh times the prevailing natural gas price (a level that should exceed the marginal costs of virtually all of the generators in ERCOT).

As the figure shows, the number of price spikes increased sharply after August 2004. There was an average of 23 price spike intervals per month during the first eight months of 2004. The number of price spike intervals more than quadrupled to 112 per month during the subsequent sixteen months. To measure the impact of these price spikes on average price levels, the figure also shows the average prices with and without the price spike intervals. The top portions of the stacked bars show the impact of price spikes on monthly average price levels. The impact grows with the frequency of the price spikes, averaging approximately $1 per MWh during the first eight months and more than $6 per MWh during the latter period. Even though price spikes account for a small portion of the total intervals, they have a significant impact on overall price levels.

Figure 5: Average Balancing Energy Prices and Number of Price Spikes 2004 & 2005



Price spikes in the markets for ancillary services have also risen significantly over this period. During the first eight months of 2004, there were 13 price spike hours for regulation up, 2 for regulation down, and 1 for responsive reserves. However, from September 2004 through April 2005, the number of price spike hours rose dramatically to 25 per month for regulation up, 34 per month for regulation down, and 20 per month for responsive reserves.[10] Since the same resources are used to supply ancillary services and energy, increases in energy prices should lead to corresponding increases in ancillary services prices. The relationship between balancing energy prices and ancillary services prices is discussed in greater detail later in this section.

While the price spikes directly impact a small portion of the total consumption of energy and ancillary services, persistent price spikes will eventually flow through to consumers. The price spikes have generally become more frequent and have become a larger component of the average balancing energy prices. There are several factors that have contributed to the rise in price spikes that are analyzed in detail in subsequent sections of this report. To the extent that price

[10] Price spikes are defined as hours where the price exceeds a threshold of $50 per MW for regulation up, regulation down, and responsive reserves.

spikes reflect true scarcity of generation resources, they send efficient economic signals in the short-run for commitment and dispatch, and in the long-run for new investment. However, to the extent that price spikes occur when economic resources are not efficiently utilized, it raises costs to consumers and sends inefficient economic signals.

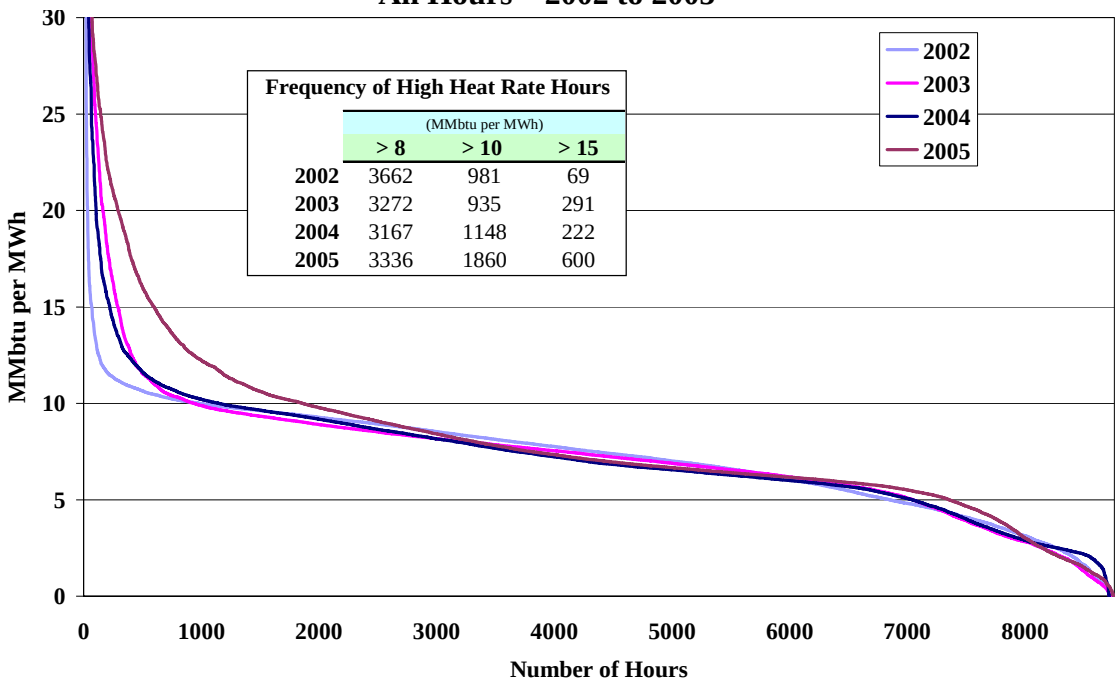
2. Balancing Energy Prices Adjusted for Fuel Price Changes

The pricing patterns shown in the prior sub-section are driven to a large extent by changes in fuel prices, natural gas prices in particular. However, prices are influenced by a number of other factors as well. To clearly identify changes in electricity prices that are not driven by changes in natural gas prices, the following figure includes two charts showing balancing energy prices corrected for natural gas price fluctuations. The first chart shows a duration curve where the balancing energy price is replaced by the marginal heat rate that would be implied if natural gas were always on the margin. The *Implied Marginal Heat Rate* equals the *Balancing Energy Price* divided by the *Natural Gas Price*.[11] The second chart shows the same duration curves for the top five percent of hours in each year. The figure shows duration curves for the implied marginal heat rate for 2002 to 2005.

In contrast to Figure 4 above, Figure 6 shows that the implied marginal heat rates were relatively consistent across the majority of hours from 2002 to 2005. For instance, the table in Figure 6 indicates that the number of hours when the implied heat rate exceeded 8 MMbtu per MWh was relatively consistent across the four years. The rise in energy prices from 2002 to 2005 is much less dramatic when we explicitly control for fuel price changes, which confirms that the increase in prices in most hours is primarily due to the rise in natural gas prices. However, the price differences that were apparent from Figure 4 in the highest-priced hours persist even after the adjustment for natural gas prices. This clearly indicates that these differences are due to factors other than changes in natural gas prices.

[11] This methodology implicitly assumes that electricity prices move in direct proportion to changes in natural gas prices.

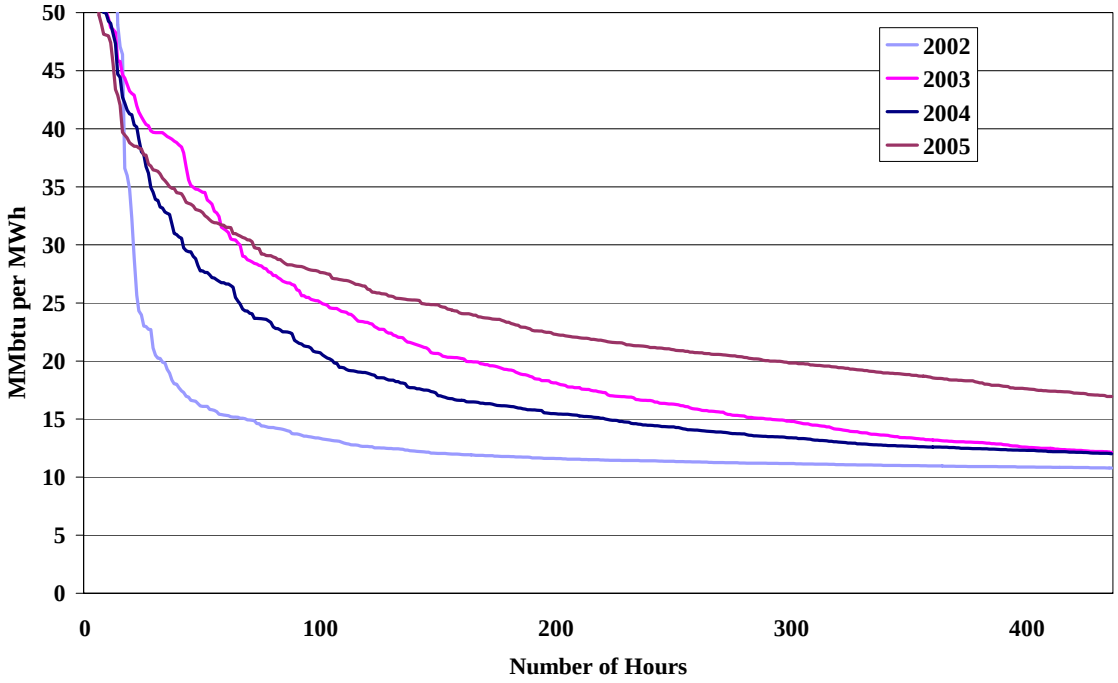
Figure 6: Implied Marginal Heat Rate Duration Curve
All Hours – 2002 to 2005



Frequency of High Heat Rate Hours

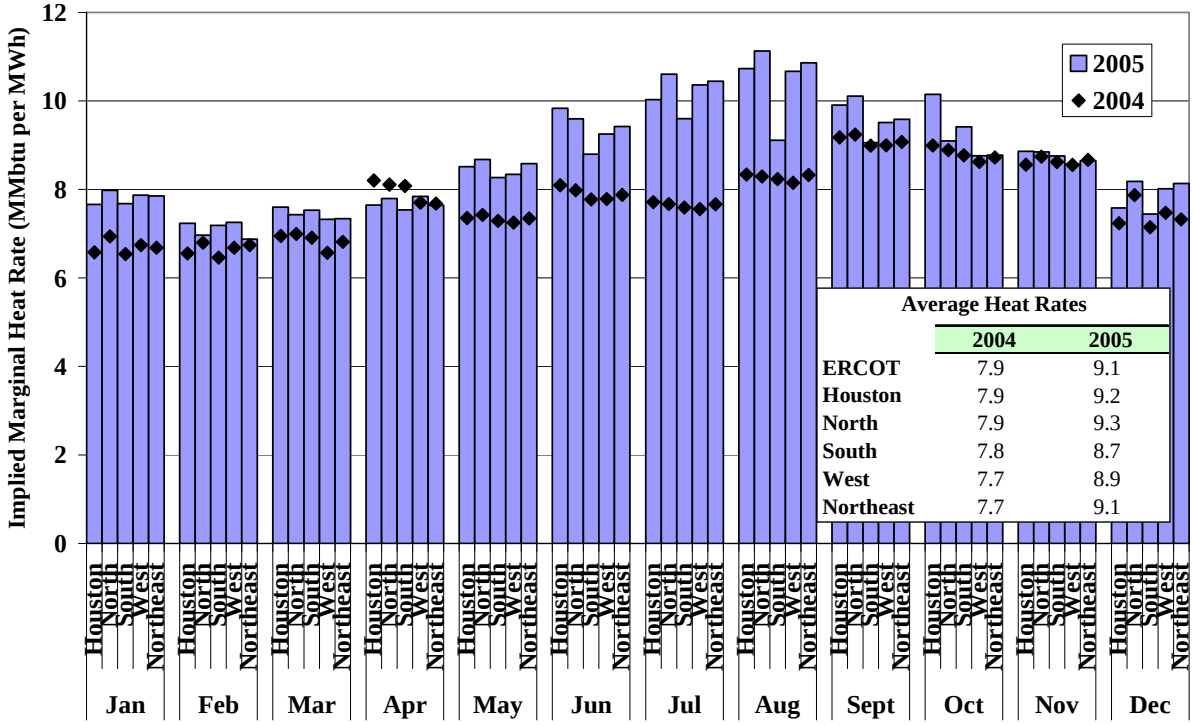
	(MMbtu per MWh)		
	> 8	> 10	> 15
2002	3662	981	69
2003	3272	935	291
2004	3167	1148	222
2005	3336	1860	600

Top Five Percent of Hours in Each Year – 2002 to 2005



To better understand these differences, the next figure shows the implied marginal heat rates on a monthly basis in each of the ERCOT zones from 2004 to 2005. This figure is the fuel price-adjusted version of Figure 1 in the prior sub-section.

Figure 7: Monthly Average Implied Marginal Heat Rates 2004 & 2005



The figure above indicates that balancing energy prices were significantly higher in 2005, even after adjusting for variations in natural gas prices. The average implied marginal heat rate rose from 7.9 MMbtu per MWh to 9.1 MMbtu per MWh, driven primarily by the rise in prices during the summer months. In 2005, August had the highest average implied marginal heat rates, averaging 10.4 MMbtu per MWh across the five zones. This was considerably higher than in August 2004, when marginal heat rates averaged slightly above 8 MMbtu per MWh. The higher average marginal heat rates in 2005 can be attributed to higher peak demand levels, less on-line capacity, and higher balancing offer prices from one supplier.

In 2004, the highest average marginal heat rates occurred in the fall. This is consistent with the summary of price spike activity in the previous sub-section which showed a substantial rise beginning in September 2004, and continuing into 2005. In a previous report, we attributed higher prices during the fall of 2004, in part, to offer patterns by TXU in the balancing energy

market. We identified 95 intervals between October 27 and December 8 when their offer patterns contributed to prices that exceeded $200/MWh.[12] If these intervals were excluded, prices would have been 8.6 percent lower from October through December 2004.

3. Price Convergence

One indicator of market performance is the extent to which forward and real-time spot 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. Relaxed balanced schedules allow QSEs to increase and decrease their purchases in the balancing energy market. This flexibility should better enable them 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. However, volatility in balancing energy prices can create risks that affect convergence between forward prices and balancing energy prices. For example, risk-averse buyers will be willing to pay a premium to purchase energy in the bilateral market.

There are several ways to measure the degree of price convergence between forward and real-time markets. In this section, we measure two aspects of convergence. The first analysis investigates whether there are systematic differences in prices between forward markets and the

[12] **"**Investigation into the Causes for the Shortages of Energy in the ERCOT Balancing Energy Market and into the Wholesale Market Activities of TXU from October to December 2004", Potomac Economics, March 2005

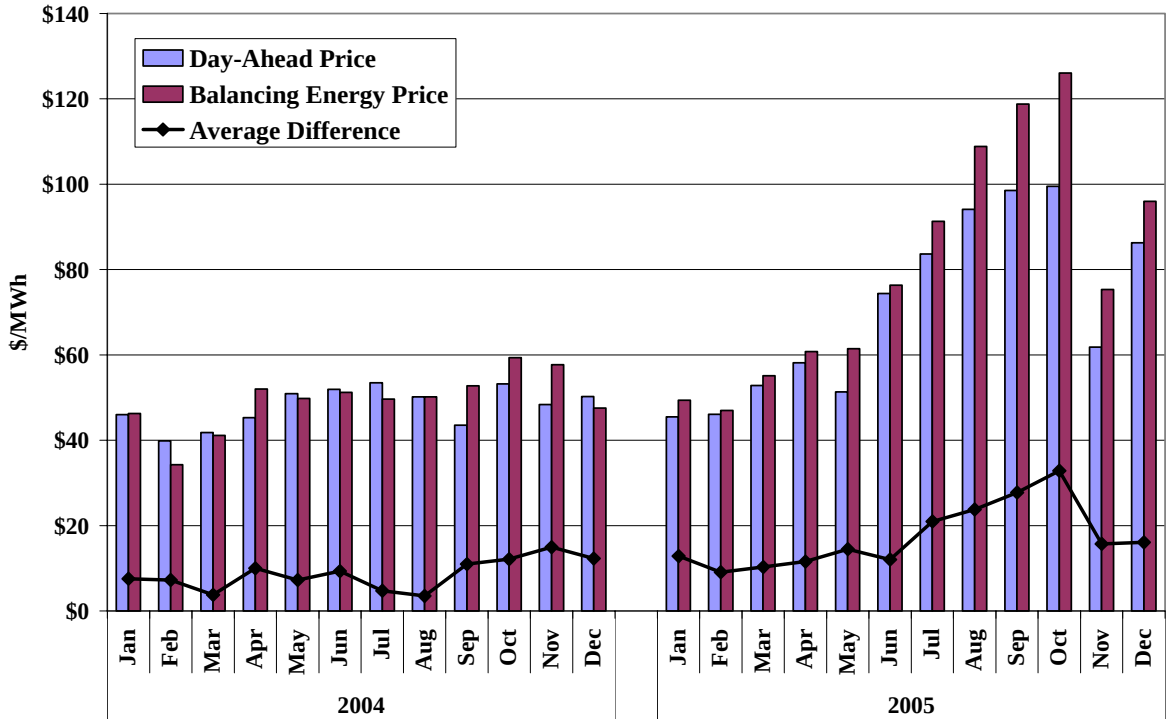
real-time market. The second tests whether there is a large spread between real-time and forward prices on a daily basis.

To determine whether there are systematic differences between forward and real-time prices, we examine the difference between the average forward price and the average balancing energy price in each month of 2004 and 2005. This reveals 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 price on a daily basis during peak hours. It is calculated by taking the absolute value of the difference between a) the average daily peak period price from the balancing energy market (i.e., the average of the 16 peak hours during weekdays) and b) the day-ahead peak hour bilateral price. This measure indicates the volatility of the daily price differences, which may be large even if the forward and balancing energy prices are the same on average. 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 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. These two statistics are shown in Figure 8 for each month in 2004 and 2005.

Figure 8 shows price convergence during peak periods (i.e. weekdays between 6 AM and 10 PM). This timeframe matches the definition of peak hours that are commonly traded in the forward market. During most of 2004, the average day-ahead price was consistent with the average balancing energy price. However, starting in September 2004 and continuing through 2005, it became common for the average balancing energy price to exceed the day-ahead price by a significant margin. Of the 16 months during this period, December 2004 was the only month when the day-ahead price was greater than the spot price.

Figure 8: Convergence Between Forward and Real-Time Energy Prices 2004 & 2005



Although forward market prices generally converge with spot market prices, unexpected spot market events can result in large differences between forward and spot prices. These unexpected events can have a substantial impact on the difference in monthly average prices. In 2005, forward prices were comparable to spot prices on the vast majority of days. However, there were 22 days when the forward price differed from the spot price by more than $50 per MWh, and the forward price was lower than the spot price on 21 of these 22 days. In general, this indicates that market participants have had difficulty predicting price spikes in the real-time market, causing the expected value of the price spikes to not be reflected in the day-ahead prices. It is also possible that the sample of day-ahead bilateral trades summarized in Figure 8 may not be representative of the market as a whole.

Figure 8 also shows that the average absolute price difference exhibited a similar pattern. The difference (shown by the line) was relatively low during the first eight months of 2004 before rising considerably during the latter period. In 2005, the average absolute difference rose sharply in the late summer and fall. As noted above, the average absolute difference measures the volatility of the price differences. Although the rise in natural gas prices contributed to the increase in price differences in late 2005, the increase in frequency of price spikes has also

played an important role by making balancing energy market prices more volatile and more difficult to predict.

The results in this section indicate that convergence between the day-ahead bilateral prices and the balancing energy prices has not consistently improved over time. Price volatility has increased in the balancing energy market has made it more difficult for forward markets to accurately forecast spot prices.

4. Volume of Energy Traded in the Balancing Energy Market

In addition to signaling 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 energy 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 energy 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 the prices obtained by 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 energy 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 9 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 9: Average Quantities Cleared in the Balancing Energy Market 2002 to 2005

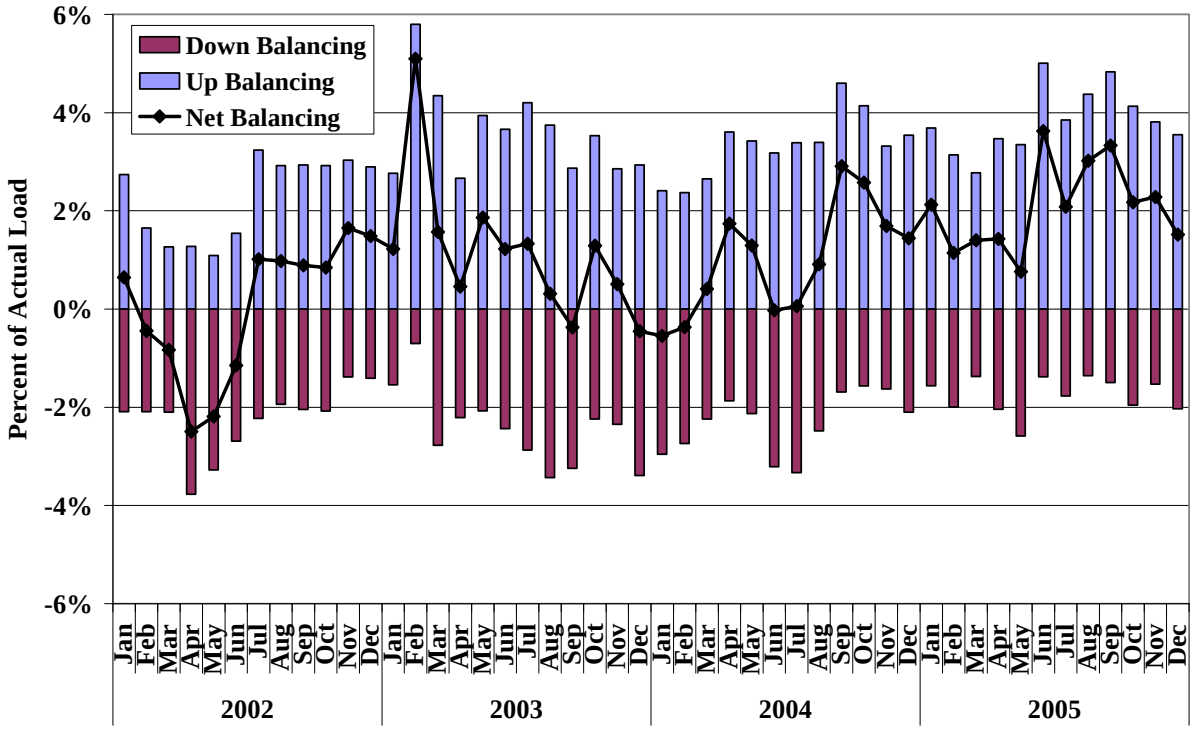


Figure 9 shows that the total volume of balancing up and balancing down energy as a share of actual load increased from an average of 4.6 percent in 2002 to 6.1 percent in 2003, 5.7 percent in 2004, and 5.6 percent in 2005. Thus, there was a general increase in trading through the balancing energy market after 2002. Over time, the volume of balancing up energy has risen relative to the volume of balancing down energy. In 2005, the average amount of net balancing up energy (i.e. balancing up minus balancing down) was 2.2 percent. Hence, market participants

generally schedule less than their full load and rely on the balancing energy market to satisfy the remaining unscheduled load.

Relaxed balanced schedules allow 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 operate as a centralized energy spot market. Although convergence between forward prices and spot prices has not been good on a consistent basis, the centralized nature of the spot market facilitates participation in the spot market and improves the efficiency of the market results.

Aside from the introduction of relaxed balanced schedules, another reason the balancing energy quantities increased after 2002 was that large quantities of balancing up and balancing down energy are deployed simultaneously to clear "overlapping" balancing energy offers. Deployment of overlapping offers improves efficiency because it displaces higher-cost energy with lower-cost energy, lowering the overall costs of serving load and allowing the balancing energy price to more accurately reflect the marginal value of energy.

When large quantities of net balancing-up or net balancing-down energy are scheduled, it indicates that Qualified Scheduling Entities (QSEs) are systematically under-scheduling or over-scheduling load relative to real-time needs. If large hourly under-scheduling or over-scheduling occurs suddenly, the balancing energy market can lack the ramping capability (i.e., how quickly on-line generation can increase or decrease its output) 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. Indeed, the tendency toward net up balancing energy purchases outside the summer helps to explain the prevalence of price spikes during off-peak months. The remainder of this sub-section and the next section will examine in detail the patterns of over-scheduling and under-scheduling that has 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 10 presents a distribution of the hourly net balancing energy. The distribution is shown on an hourly basis rather than by interval

to minimize the effect of short-term ramp constraints and to highlight the market impact of persistent under- and over-scheduling. Each of the bars in Figure 10 shows the portion of the hours during 2005 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 on average) in approximately 16 percent of the hours in 2005.

Figure 10: Magnitude of Net Balancing Energy and Corresponding Price 2005

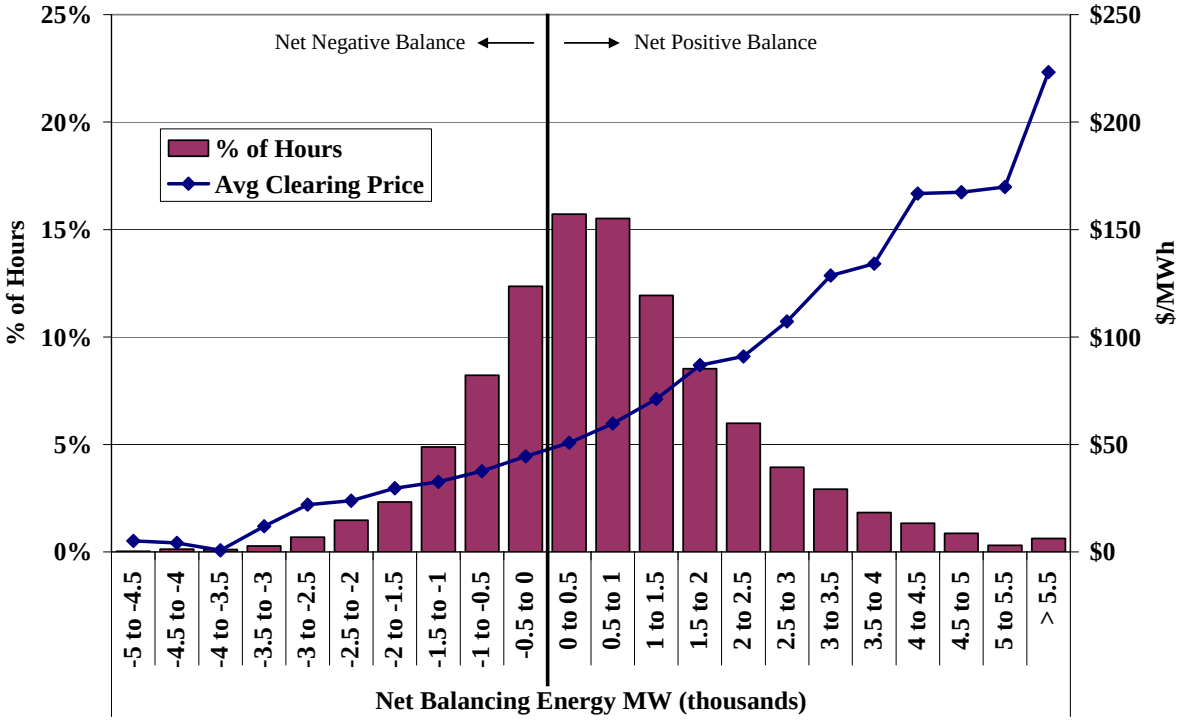


Figure 10 shows a relatively symmetrical distribution of net balancing energy purchases centered around 0.5 gigawatts. This is consistent with Figure 9 which showed that there were substantial net balancing up quantities on average in each month during 2005. In approximately 52 percent of the hourly observations shown, Figure 10 also shows that net balancing energy schedules averaged between -1.0 and 1.0 gigawatts.[13] Hence, there were many hours when the net balancing energy traded was relatively low, because the total scheduled energy is frequently close to the actual load.

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

The line plotted in Figure 10 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 clearly indicates that balancing energy prices increase as net balancing energy volumes increase. This is also consistent with the patterns of prices and volumes in 2003 and 2004.[14] The pattern indicates that the balancing energy market is thinly traded, which can undermine its efficiency. We analyze this relationship more closely in the next sub-section, and in Section II we discuss how scheduling practices and ramping issues explain much of the observed pattern.

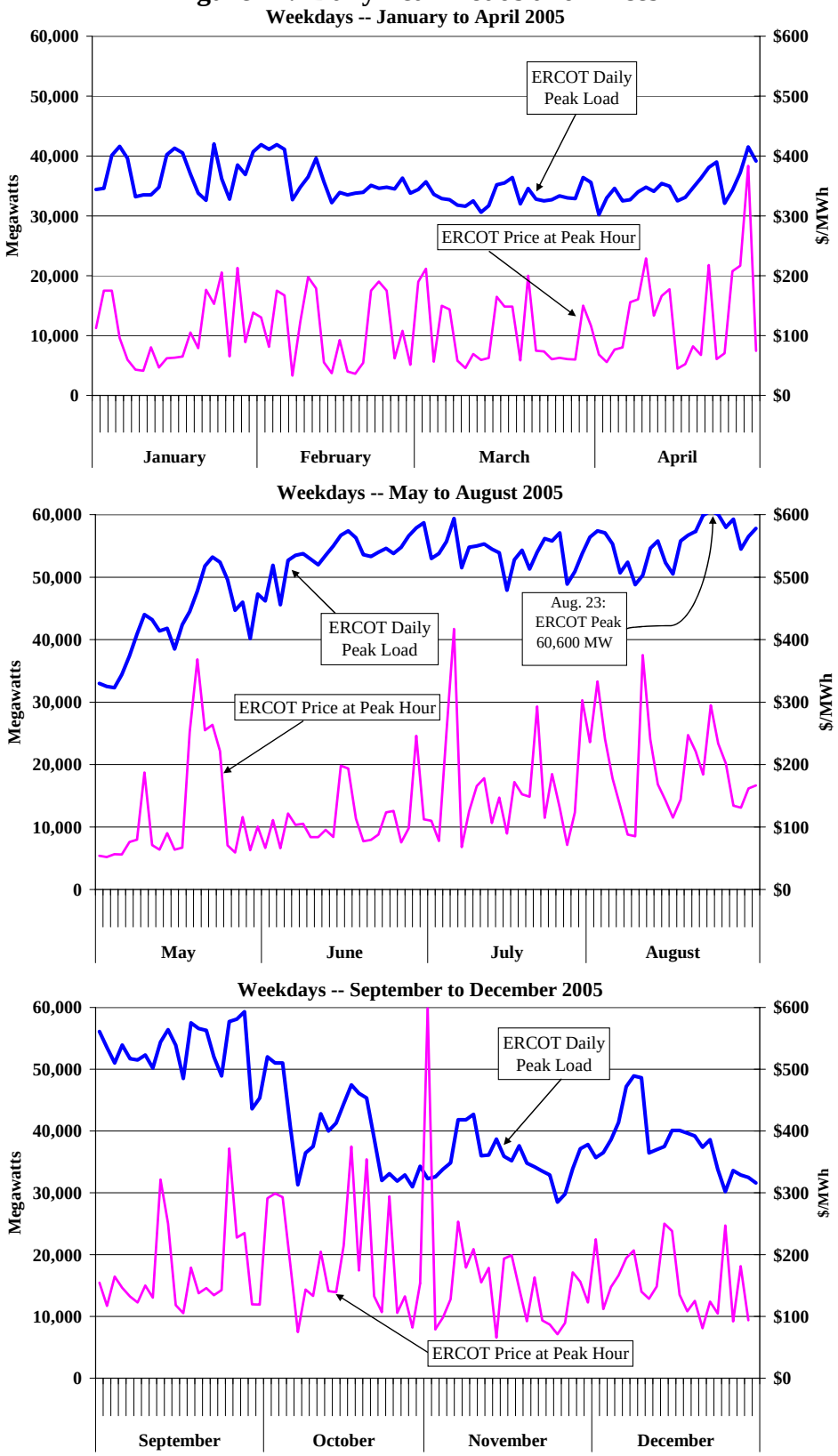
5. Determinants of Balancing Energy Prices

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 11 shows the average balancing energy price and the actual load in the peak hour of each weekday during 2005.

[14] See 2003 SOM Report and 2004 SOM Report

Figure 11: Daily Peak Loads and Prices

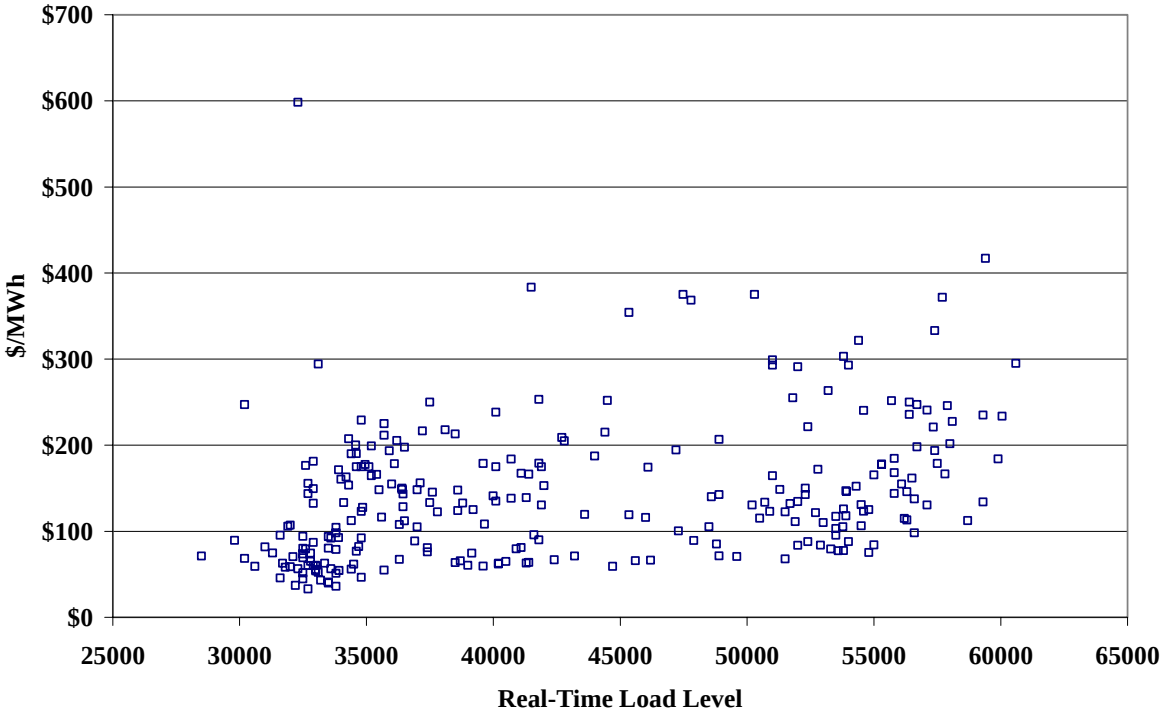


The figure shows that a large share of the days with high prices (e.g., greater than $200/MWh) coincide with periods when demand is high or rising quickly relative to the previous several days. However, prices spikes also occurred during lower demand periods. The price averaged $295 per MWh during the highest load hour on record, which occurred on August 23, 2005. Although this is quite high, there were 12 days in 2005 when prices exceeded this level in the peak demand hour of the day.

In an efficient market, we expect for peak prices to occur under extreme demand conditions or as a result of unforeseen conditions that cause brief shortages, such as the loss of a large generator or a sudden rise in load. In ERCOT, prices in the balancing market can reach extremely high levels even when demand is not particularly high. In Section II, we illustrate that substantial amounts of on-line and quick start capacity are routinely not offered to the balancing energy market. This can lead to price spikes and shortages in the balancing market when excess capacity is available.

To further examine the relationship between actual load in ERCOT and the balancing energy prices, Figure 12 shows the same data as Figure 11, but plots the average balancing energy prices versus the daily peak loads in ERCOT irrespective of time. This type of analysis shows more directly the relationship between balancing energy prices and actual load. In a well-performing market, one should expect a clear positive relationship between these variables since resources with higher marginal costs must be dispatched to serve rising load.

Figure 12: ERCOT Balancing Energy Price vs. Real-Time Load
Weekdays -- Peak Load Hour -- 2005



The figure indicates a positive correlation between real-time load and the clearing price in the balancing market, although it is relatively weak. It should be a rare occurrence for the average price in the peak hour to exceed $150 per MWh when ERCOT load is less than 40 GW. However, this occurred on 28 percent of days shown in the figure, and 27 percent of the days before the period when gas prices rose to high levels. Although prices were generally higher at higher load levels, there were other factors that were more closely correlated with clearing prices. For instance, the analysis shown above in Figure 10 indicates that the net volume of energy purchased in the balancing energy market is a much stronger determinant of price spikes than the level of demand.

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 27 GW at 4 AM to 38 GW at 1 PM. Thus, the change in load averages 1,280 MW per hour (320 MW per 15-minute interval) during the morning and early afternoon. Figure 13 shows the average load and balancing energy price

in each interval from 4 AM through 1 PM in 2005. The price is plotted as a line in the figure while the average load is shown with vertical bars.

Figure 13: Average Clearing Price and Load by Time of Day
Ramping-Up Hours – 2005

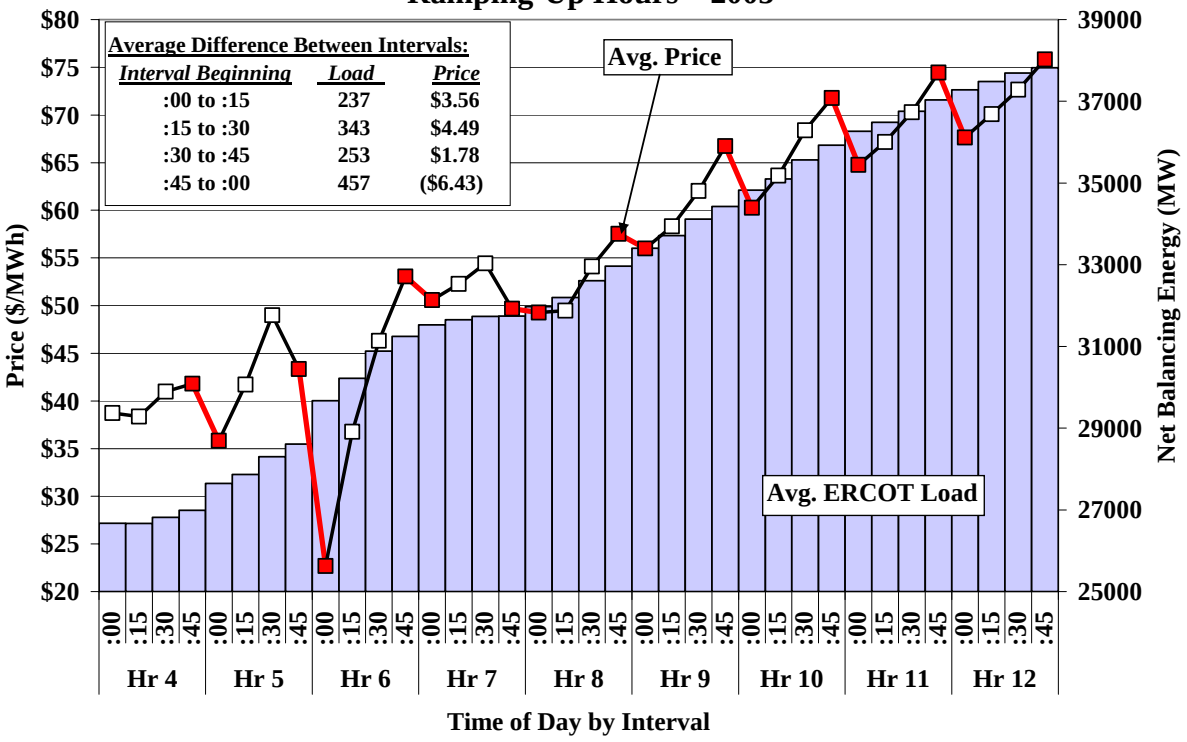


Figure 13 shows that the load steadily increases in every interval and prices generally move upward from about $39 per MWh at 4:00 AM to $76 per MWh at 12:45 PM. If actual load were the primary determinant of energy prices, the balancing energy prices would rise gradually as the actual load rises. However, Figure 13 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, the red lines highlight 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 -$6.43 per MWh. This occurs because participants tend to change their schedules once per hour, bringing on additional substantial quantities of generation at the beginning of the hour that reduces the balancing energy prices.

A similar pattern is observed at the end of the day when load is decreasing. In ERCOT, load tends to decrease 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 1,840 MW per hour (460 MW

per 15-minute interval) during the late evening. Figure 14 shows this decrease in load by interval, together with the average balancing energy prices for the intervals from 9 PM to 3 AM.

Figure 14: Average Clearing Price and Load by Time of Day
Ramping-Down Hours – 2005

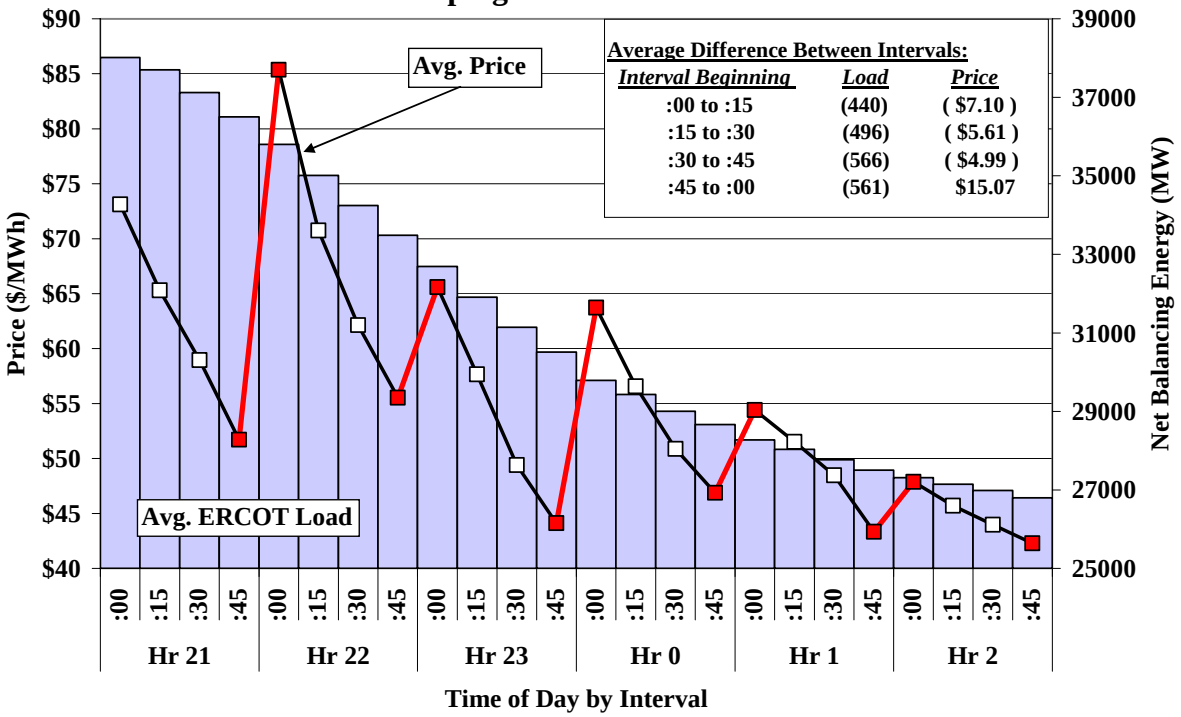


Figure 14 shows that while balancing energy prices decrease over these intervals, they follow a similar pattern as exhibited in the ramping-up hours. The balancing energy price decreases in each interval of the hour before rising substantially in the first interval of the following hour. The balancing energy price increases by an average of $15.07 per MWh from the last interval of one hour to the first interval of the next hour during this period. This occurs because participants tend to change their schedules once per hour, de-committing generating resources at the beginning of the hour. Because the supply decreases at the beginning of these hours by much more than load decreases, the balancing energy prices generally increase. This is consistent with the patterns of energy schedules and balancing prices in 2003 and 2004.[15]

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,

15 See 2003 SOM Report and 2004 SOM Report

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.

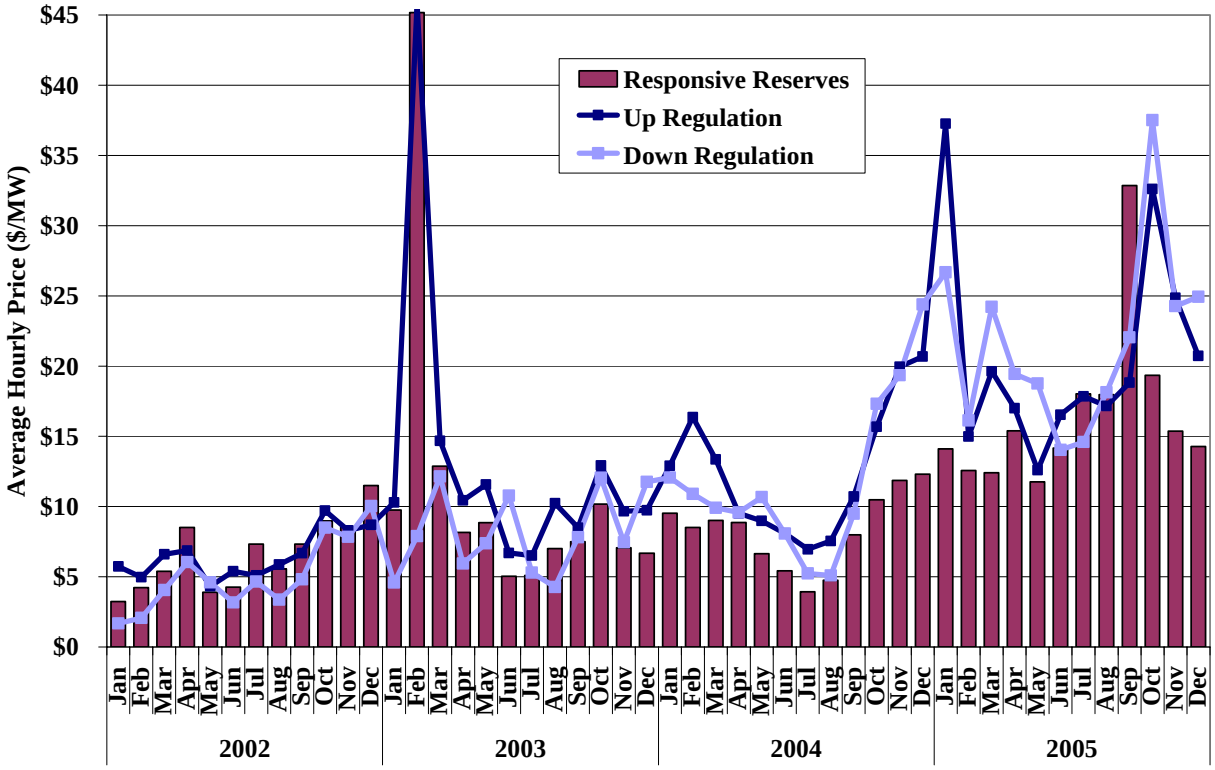
B. Ancillary Services Market Results

The 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 ancillary services markets in 2005.

1. Reserves and Regulation Prices

Our first analysis in this section provides a summary of the ancillary services prices over the past four years. Figure 15 shows the monthly average ancillary services prices between 2002 and 2005. Average prices for each ancillary service are weighted by the quantities required in each hour.

Figure 15: Monthly Average Ancillary Service Prices 2002 to 2005



This figure shows that ancillary services prices have generally risen over the last four years. Much of this increase can be attributed to the increase in energy prices that occurred over the same timeframe. Because ancillary services 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 below the most profitable level. From 2002 through 2004, regulation down prices were lower than regulation up prices, indicating that the opportunity costs were greater for providers of regulation up. However, over time, the pattern has shifted so that regulation down prices were 4 percent higher on average than regulation up prices in 2005. This pattern and the factors that explain it are discussed in greater detail later in this section following Figure 19.

The figure also shows that the prices for up regulation generally exceed prices for responsive reserves. This is consistent with expectations because a supplier must incur 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. However, during periods of persistent high prices, regulation up providers may have lower opportunity costs than responsive reserves providers to the extent that they are dispatched up to provide regulation. This may explain why responsive reserves prices were higher on a monthly average basis than regulation up prices from July through September 2005.

Figure 15 also shows that reserves and regulation prices vary according to the season. In ERCOT before 2005, ancillary services prices have been lower during the summer than at other times of the year. This may be attributable to the fact that the required quantities of reserves and regulation are relatively constant over the year while the supply of resources that can provide reserves and regulation (i.e., on-line capacity not scheduled for energy) tends to increase in proportion to load. The additional supply puts downward pressure on reserves and regulation

prices. However, ancillary services prices can also rise considerably when capacity becomes scarce under peak demand conditions during the summer and winter. Energy prices rose more notably during the summer of 2005 than in previous summers, resulting in a rise in responsive reserves prices relative to non-summer months. In the other wholesale markets in the U.S., where reserve margins are generally lower than in ERCOT, responsive reserves prices also tend to rise during the summer months.

Beginning in September 2004 and continuing through 2005, ancillary services prices have been consistently higher than at any other time since the wholesale market began operation, with the exception of February 2003. This has occurred for two reasons. First, the frequency of price spikes in the balancing energy market has been substantially higher during this period, and this has raised the opportunity costs of providing ancillary services. The frequency of price spikes is shown in Figure 5.

Second, there was a reduction in the average quantity of on-line capacity that was committed and not scheduled for energy that coincided with the rise in prices in the fall of 2004.[16] This may indicate that the market is meeting load and ancillary services obligations more efficiently during this period than in prior periods. However, the reduction in excess committed capacity increases the opportunity costs of ancillary services providers by increasing the likelihood that their on-line capacity will be needed for balancing energy.

In October 2005, ERCOT modified the day-ahead procurement process for ancillary services so the markets for regulation, responsive reserves, and non-spinning reserves clear simultaneously. By running a simultaneous auction for the four services, it is possible to clear the market with the least expensive set of available offers. This change is likely to result in more efficient prices for ancillary services since they will better reflect the opportunity costs of not providing the other services. In the longer-run, co-optimizing the procurement of these services with energy in the real-time market would provide substantial additional benefits for the ERCOT market.

One way to evaluate the rationality of prices in the ancillary services markets is to compare the prices for different services to determine whether they exhibit a pattern that is reasonable relative

[16] See Figure 62, 2004 SOM Report

to each other. Table 1 shows such an analysis, comparing the average prices for responsive reserves and non-spinning reserves over the past four years in those hours when ERCOT procured non-spinning reserves. Non-spinning reserves were purchased in approximately 18 percent of the hours during 2002, 25 percent of hours during 2003, 24 percent of hours during 2004, and 23 percent of hours during 2005.

Table 1: **Average Hourly Responsive Reserves and Non-Spinning Reserves Prices During Hours When Non-Spinning Reserves Were Procured 2002 to 2005**

	2002	2003	2004	2005
Non-Spinning Reserve Price	$14.51	$9.85	$6.83	$25.10
Responsive Reserve Price	$9.20	$10.73	$9.10	$28.16

Table 1 shows that responsive reserves prices are higher on average than non-spinning reserves prices.[17] The prices in 2002 were the exception because non-spinning reserves prices were above $990 per MWh for 13 hours on two days. It is reasonable that responsive reserves prices would generally be higher since responsive reserves are a higher quality product which that be delivered in 10 minutes from on-line resources while non-spinning reserves must be delivered in 30 minutes. Thus, it might seem counterintuitive that non-spinning reserves prices were higher than responsive reserves prices in 18 percent of hours during 2002, 35 percent of hours during 2003, 37 percent of hours during 2004, and 38 percent of hours during 2005. However, ERCOT uses non-spinning reserves and responsive reserves differently, which can cause non-spinning reserves to be more costly for some resources to provide than responsive reserves.

Generators incur two types of costs associated with providing reserves in the ERCOT market. First, reserves providers incur opportunity costs from any profitable sales they forego in the energy market. For generators, this is the same regardless of whether the generator is providing responsive or non-spinning reserves. The second cost that must be considered is the cost of actually being called upon by ERCOT to deploy reserves in real-time. Since generators deployed for reserves are paid for the resulting output at the balancing energy price, there is a risk of being deployed when the balancing energy price is lower than the generator's production

[17] The values in this table differ slightly from the values found in the 2004 report due to a change in the weighting of the prices.

costs. While it is also possible for the generator to benefit when the balancing energy price is higher than the generator's costs, this occurs less frequently. Thus, generators providing reserves generally run at a loss when they are deployed by ERCOT.

The expected costs of being deployed for reserves are based on the following two factors: (a) the average difference between the resource's production cost and the balancing energy price, and (b) the probability of being deployed. In 2005, less than 0.1 percent of the responsive reserves were actually deployed, while 4.3 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.

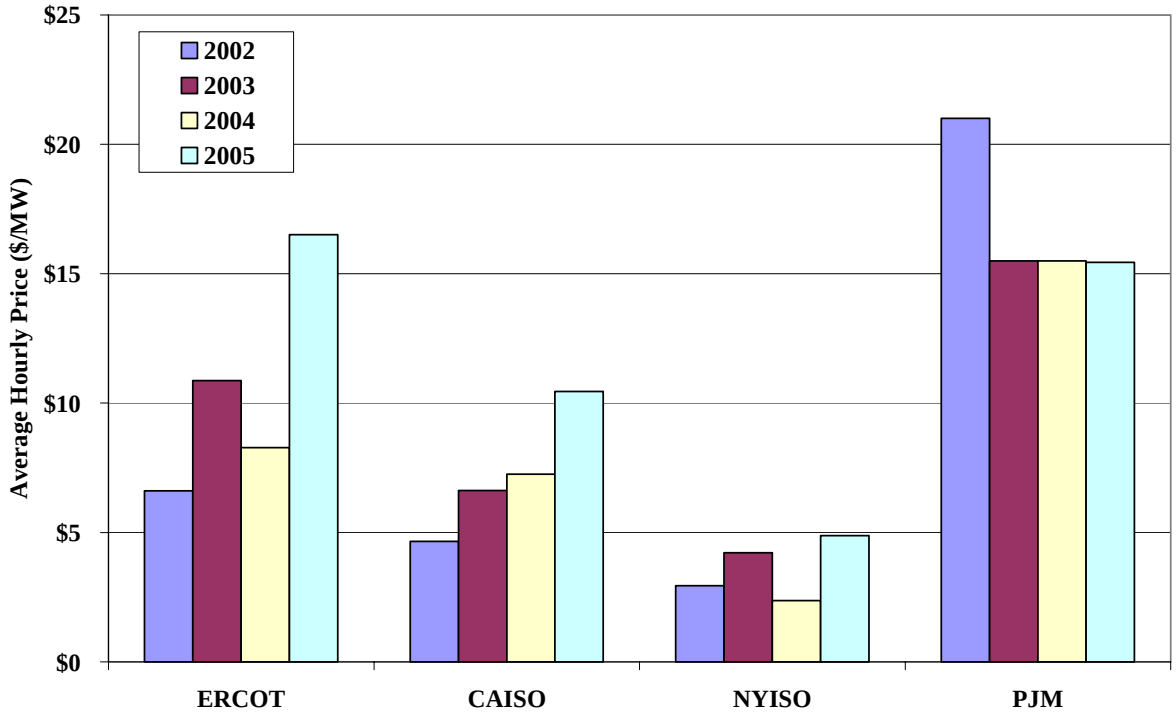
In general, the purpose of operating reserves is to protect the system against unforeseen contingencies (e.g., transmission line or generator outages), rather than for meeting load. The balancing energy market deployments that occur in the 15-minute timeframe and regulation deployments that occur in the 4-second timeframe are the primary means for meeting the load requirements. However, in cases when the resources in the balancing energy 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 the long-run, after the implementation of the nodal market design, we recommend that ancillary services and energy be jointly optimized in the real-time market. One benefit of co-optimization is that it can help improve the accuracy of price signals during non-spinning reserve deployments. Currently, ERCOT deploys non-spinning reserves manually when it is in danger of running short of energy. During these periods of shortage, the additional supply that is added to the energy market has the effect of dampening scarcity price signals. However, in a market where energy and reserves are co-optimized, the deployment of reserves for energy needs will be done by the market based on the economic trade-offs between reserves and energy. This

approach results in clearing prices for energy, responsive reserves and non-spinning reserves that efficiently reflect the shortage conditions.

Responsive reserves prices rose considerably in 2005, averaging around $16 per MWh for the year. Figure 16 shows how the annual average prices in ERCOT from 2002 to 2005 compare to the responsive reserve prices in the California, PJM, and New York wholesale markets. The figure shows that the responsive reserve prices in ERCOT were higher than comparable prices in California, New York, and PJM during 2005.

Figure 16: Responsive Reserves Prices in Other RTO Markets 2002 to 2005



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, nearly one half of ERCOT's responsive reserves are satisfied by demand-side resources offered at very low prices, which should serve to offset the fact that ERCOT procures a higher quantity of responsive reserves.

A second reason ERCOT Responsive Reserve prices are higher is because ERCOT (like California and PJM) does not jointly-optimize ancillary services and energy markets. 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 needs substantially more regulating capacity during ramping hours. When demand rises, higher-cost resources must be employed and prices should increase.

Figure 17 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.

The figure shows that ERCOT requires approximately 1,100 MW of capability prior to the initial ramping period (beginning at 6 AM). The requirement then jumps up to 2,000 MW during the steepest ramping hours from 6 AM to 9 AM. The requirement declines to 1,400 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 1,950 MW.

Figure 17: Regulation Prices and Requirements by Hour of Day 2005

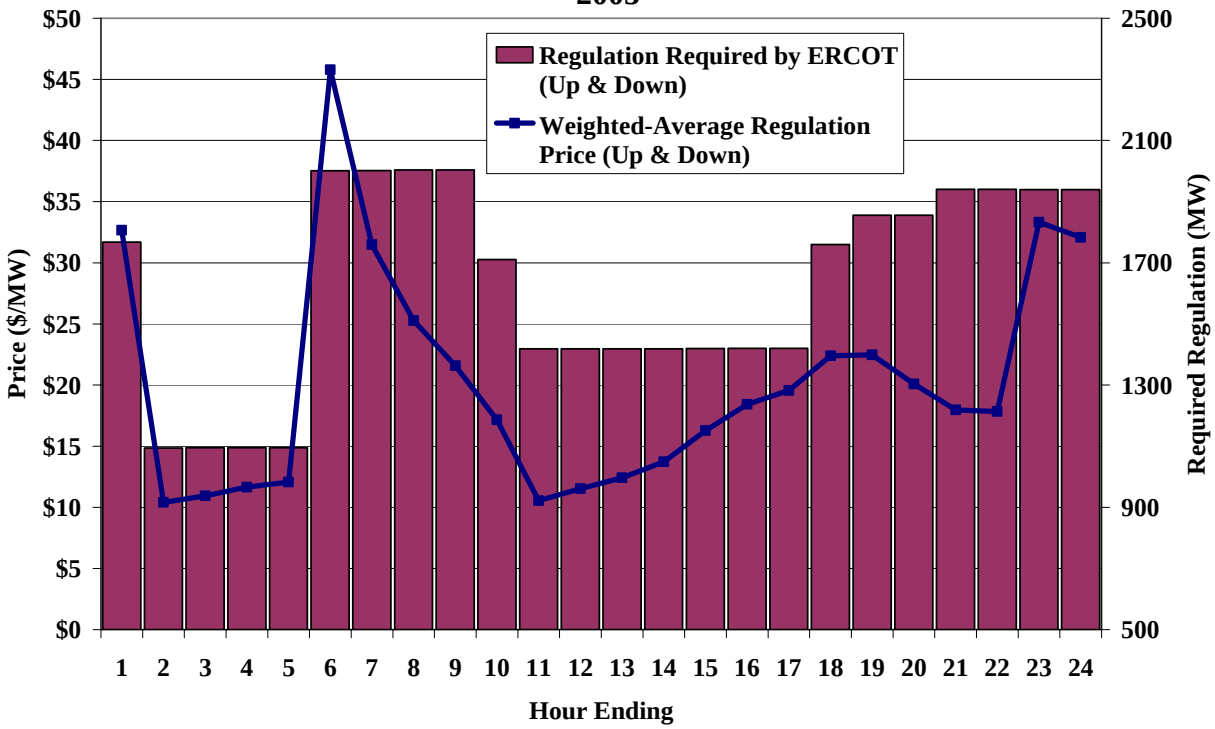


Figure 17 indicates that average regulation prices are closely correlated with the regulation quantity purchased. During non-ramping hours, such as overnight, late morning, and in the afternoon, regulation prices average from $10 to $20 per MW. During the ramping hours in early morning and evening, average regulation prices range from $17 to $46 per MW. The higher prices during ramping hours also occur because a larger portion of regulation capability is actually deployed during ramping hours. Additionally, the price range exhibited in the ramping hours was wider in 2005 than in the previous years. This is largely due to the higher overall regulation prices that have occurred since late in 2004, when lower levels of excess online generating capacity and increased balancing energy market volatility resulted in higher and more volatile regulation prices.

Regulation prices are particularly high during hours ending 1, 6, 7, 23, and 24. Less supply is available during these hours, because many regulation-capable units in ERCOT start after 7 am and shutdown before 10 pm. This reduces the amount of capacity available to supply regulation, which leads to higher prices.

Although regulation prices have risen markedly since 2002 due to several factors discussed above, ERCOT has taken significant steps over the same period to reduce regulation market costs. ERCOT has gradually reduced the amount of regulation it procures and uses to keep supply and demand in balance and control frequency on the system. This has directly reduced regulation costs by reducing the quantity scheduled. However, this has also indirectly reduced regulation costs by lower the clearing prices of regulation. Figure 18 summarizes the average amounts of regulation procured through the auction and/or bilateral arrangements on an annual basis since 2002.

Figure 18: Annual Average Regulation Procurement 2002 to 2005

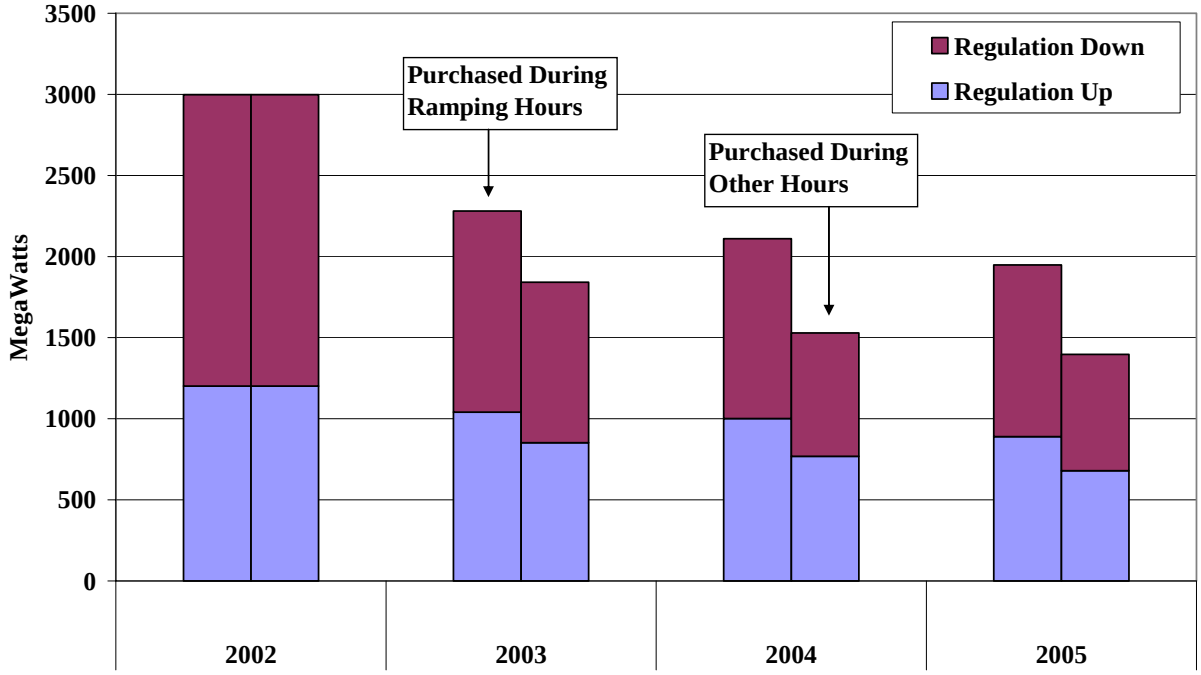
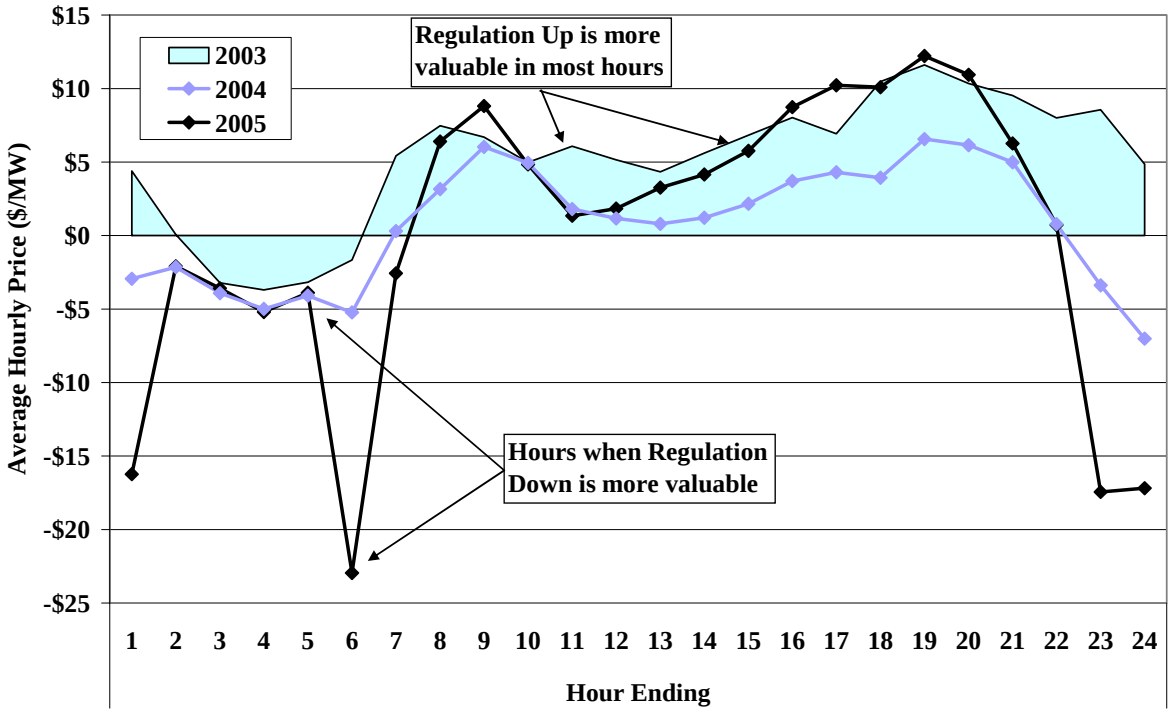


Figure 18 shows that ERCOT has reduced the average regulation quantity scheduled since 2002. The largest reduction was from 2002 to 2003, although the reductions in the remaining two years were also substantial. Overall, ERCOT has lowered the required amount by 35 percent during ramping hours and 53 percent during non-ramping hours. During the same period, ERCOT also adjusted the relative shares of regulation up and regulation down with the regulation down share decreasing from 60 percent in 2002 to close to 50 percent in 2005.

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. Like the comparative analysis of responsive reserve and non-spinning reserve prices presented earlier in this sections, our next analysis examines the differences between up and down regulation prices to determine whether they exhibit a rational pattern that is consistent with market fundamentals. Figure 19 shows the average up regulation price minus the average down regulation price in each hour of the day for 2003, 2004, and 2005 separately.

Figure 19: Comparison of Up Regulation and Down Regulation Prices
Up Regulation Price Minus Down Regulation Price
2003 to 2005



The figure reveals a distinct intertemporal variation in the regulation 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.

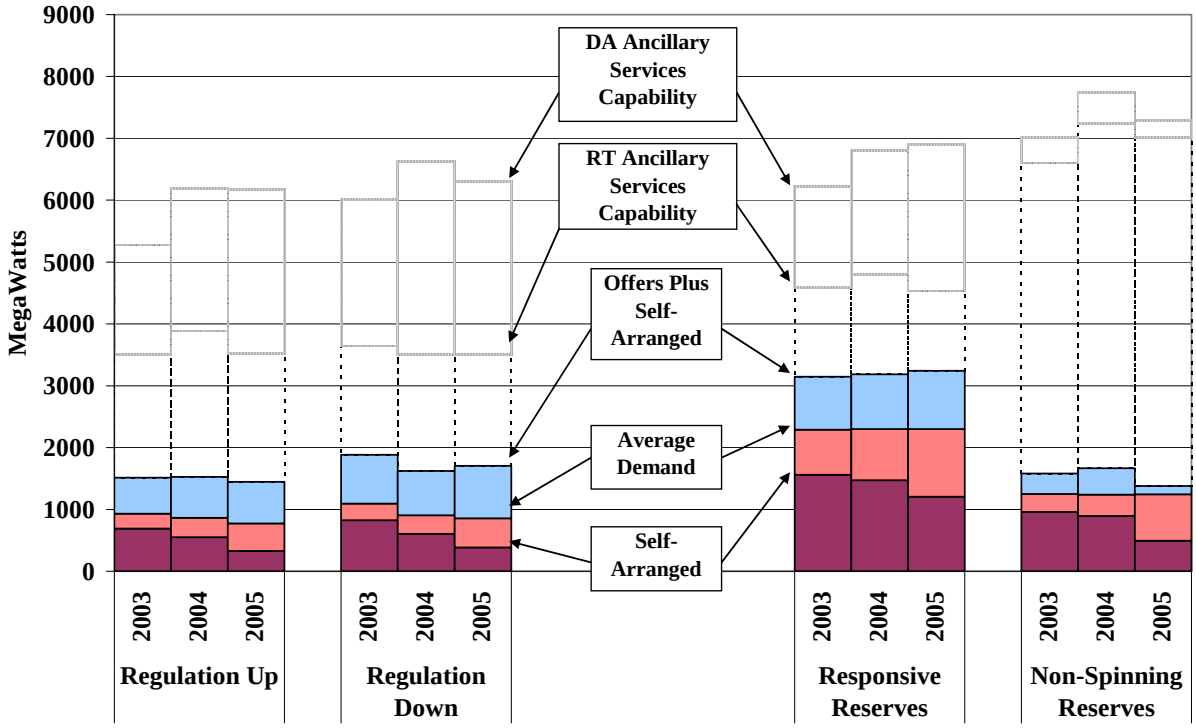
Figure 19 also shows a significant downward shift in the price difference from 2003 to 2004 and again from 2004 to 2005, which means that up regulation has become less expensive relative to down regulation. The figure also reveals that the positive and negative differences between regulation up and regulation down prices grew substantially in 2005. During the afternoon hours, the premium on regulation up has more than doubled since 2004. Conversely, the premium on regulation down increased markedly in hours-ending 1, 6, 23, and 24. Higher overall energy prices, in 2005, have lead regulation providers to incur higher opportunity costs. Naturally, this magnifies cost differences between providing up and down regulation.

2. Provision of Ancillary Services

To better understand the reserve 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 20. This figure 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 Figure 20 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).

Figure 20: Reserves and Regulation Capacity, Offers, and Schedules 2003 to 2005



Note: Non-spinning reserve capability is based on data from generator resource plans. Regulation and responsive reserves capability is based on ERCOT data.

The capability shown in Figure 20 incorporates ERCOT's requirements and restrictions for each type of service. 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 20 is not reduced to account for energy produced from each unit, which causes the capability on some resources to be overstated in some hours. Approximately 49 percent of the demand for responsive reserves was satisfied by Loads acting as Resources ("LaaRs"). LaaRs account for only 1150 MW of the responsive reserves capability shown above, because there is currently a requirement that no more than 50 percent of the 2300 MW requirement be met with LaaRs.

For non-spinning reserves, Figure 20 includes the capability of units that QSEs indicate are 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 20 shows that less than one-half of each type of ancillary services capability was offered during 2003, 2004, and 2005. 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 who must predict one day in advance whether their resources will be more valuable as energy or as ancillary services.

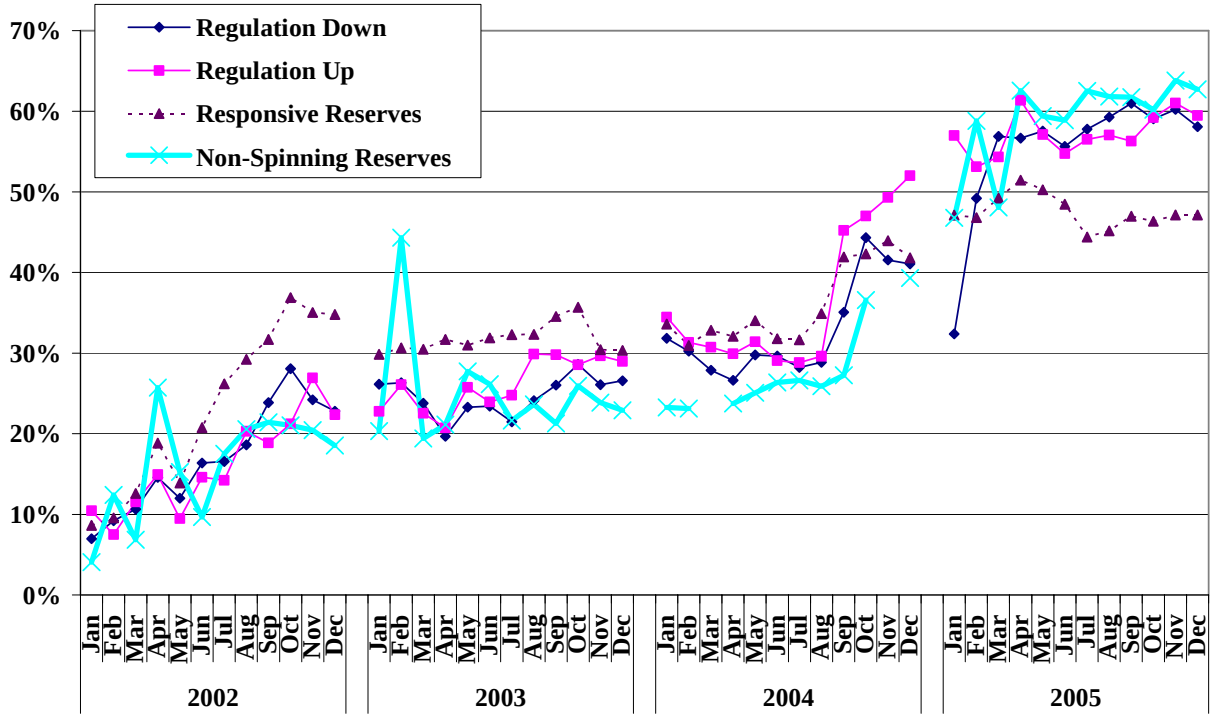
In addition, participants may not offer the capability of resources they do not expect to commit for the following day. This explanation is less likely because suppliers could submit offer prices high enough to ensure that their costs of committing additional resources to support the ancillary services offers are covered. Nonetheless, there is a substantial quantity of reserves that remain available in real time, but that are not offered. 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. However, this is not likely to be the primary reason, since both small and large participants choose not to offer substantial portions of their capability in the ancillary services market.

Figure 20 shows modest changes in the amount of day-ahead ancillary services capability between 2003 and 2005. The installation of several gigawatts of new capacity has contributed to overall capability, while the continued mothballing and retirement of certain units has reduced capability. The average amount of excess on-line capacity has declined since 2003, thereby reducing the amount of capacity available to provide ancillary services.

Finally, although market participants increasingly rely on the auction market to procure these services, Figure 21 shows that a significant share of these services is still 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 21 shows the share of each type of ancillary service that is purchased through the ERCOT market.

Figure 21: Portion of Reserves and Regulation Procured Through ERCOT 2002 to 2005



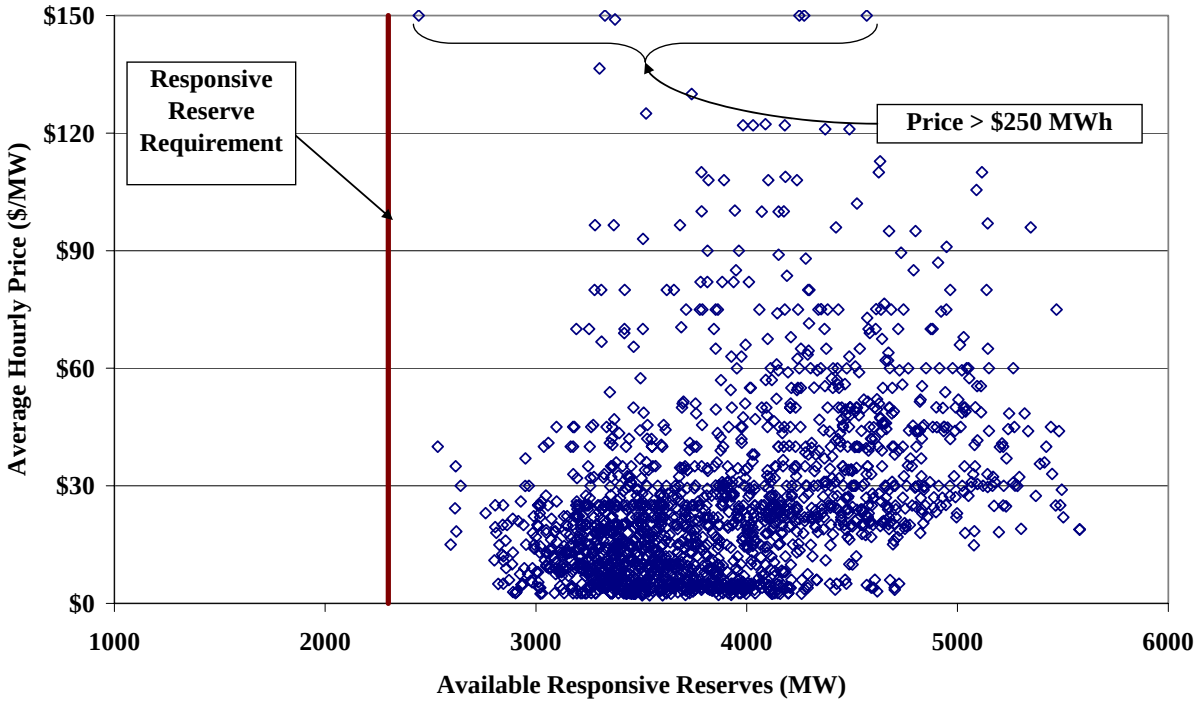
This figure shows that purchases of all ancillary services from the ERCOT markets have increased consistently over the past four years, rising sharply after the summer of 2004. As noted earlier, there was a significant rise in balancing energy prices during the fall of 2004, and since ancillary services providers must forego energy sales, this has likely increased the opportunity cost of providing ancillary services. When buyers of ancillary services face higher bilateral contract prices, it can push more of their purchases into the ERCOT market until prices between the two markets converge. As market participants have gained more experience with the ERCOT markets, larger portions of the available reserves and regulation capability have been offered into the market, thereby increasing the market's liquidity.

The next analysis in this section evaluates the prices prevailing in the responsive reserves market during 2005. 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.

Figure 22 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 and the actual dispatch point for LaaRs. 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 22: Hourly Responsive Reserves Capability vs. Market Clearing Price Afternoon Peak Hours – 2005



This figure indicates a somewhat random pattern of responsive reserves prices in relation to the hourly available responsive reserves capability in real time. In a well functioning-market for responsive reserves, we would expect excess capacity to be negatively correlated with the

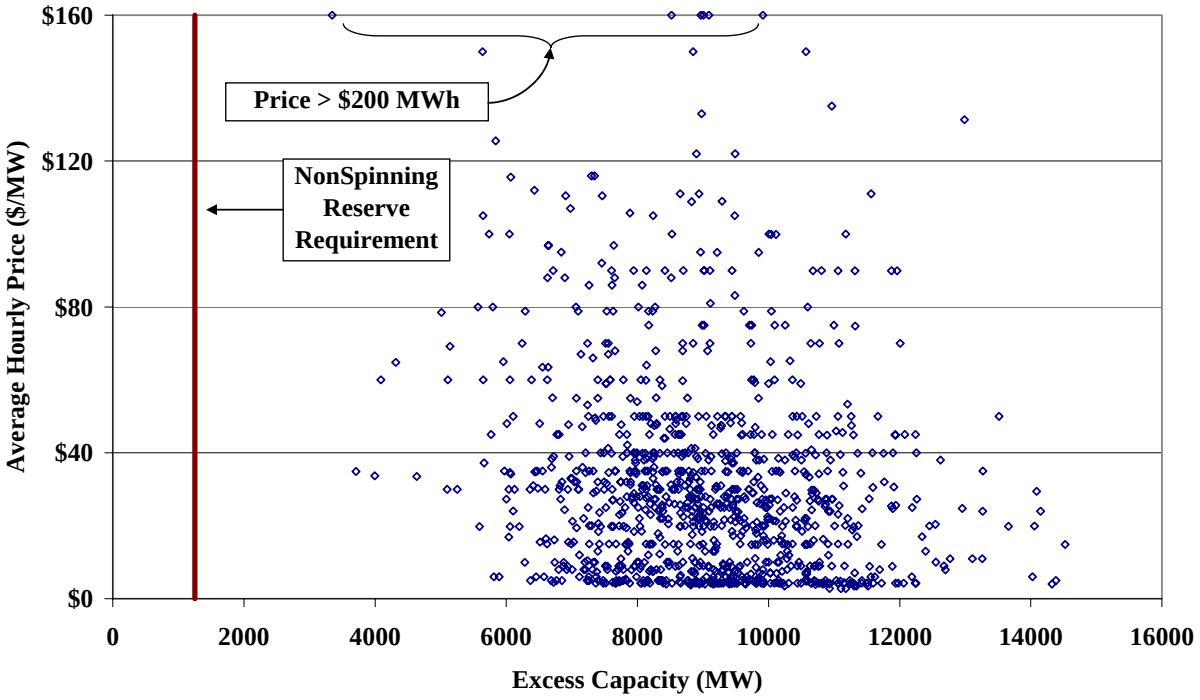
clearing prices, but this was not the case in 2005. Although the responsive reserves prices were considerably higher in 2005, similar analyses in previous reports show the same lack of correlation between prices and available reserves in 2003 and 2004. This lack of correlation raises significant issues regarding the efficiency of this market. Particularly surprising is the frequency with which the price exceeds $20 per MW 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 recommend in the context of the nodal market design which is currently being developed for implementation in 2009.

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 23 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 2005.

Like the previous analysis of responsive reserves, the results shown in Figure 23 do not indicate a significant correlation between non-spinning reserves prices and the quantity of available reserves capability in real time. This is consistent with similar analyses in previous reports which showed a lack of correlation between prices and excess capacity in 2003 and 2004.[18] In a well functioning-market for non-spinning reserves, we would expect excess capacity to be negatively correlated with the clearing prices for two reasons.

[18] The analyses mentioned in previous reports showed a lower level of excess capacity than Figure 23 because they included on-line and quick start capacity not needed for regulation or responsive reserves obligations, a more restrictive set of resources. Figure 23 includes unloaded capacity not needed for regulation or responsive reserves obligations that is flagged in the resource plan as non-spin capable.

Figure 23: Hourly Non-Spinning Reserves Capability vs. Market Clearing Price Afternoon Peak Hours – 2005



First, the opportunity cost of providing reserves rather than energy should be close to zero when substantial excess is available. Second, the probability of being deployed decreases with excess capacity. Thus, the lack of correlation between non-spinning reserves prices and excess capacity raises significant issues regarding the performance of this market. However, the lack of correlation is at least partly explained by the fact that substantial amounts of excess capacity are routinely not offered to the balancing energy market.

Although there is usually a significant excess of reserves in real-time, payments are only made to QSEs that are scheduled day-ahead (or during the adjustment period) to provide reserves. These QSEs must ensure that sufficient capacity is on-line in real-time and not scheduled for energy or another ancillary service. QSEs cannot offer energy from their responsive reserves capacity into the balancing energy market. In the event that responsive reserves are deployed, QSEs must be capable of ramping up the reserved capacity within 10 minutes. ERCOT limits that the amount of responsive reserves that can be provided by any particular resource to 20 percent or less of the resource's total capacity. This requirement is intended, in part, to ensure that responsive reserves are held on a diversified set of units and held in amounts. If one or two units providing

responsive reserves trip off-line or are congested down, it will not substantially affect the reliability of the system.

Under normal circumstances, responsive reserves are not deployed in real-time. ERCOT begins to deploy responsive reserves when frequency drops below 59.91 Hz, and a substantial share will not be deployed unless frequency drops to 59.7 Hz. These thresholds result in very infrequent real-time deployments of responsive reserves, which makes it difficult to evaluate the response of QSEs to reserve deployments. However, it is possible to evaluate whether the QSE has the capability to respond. The following analysis evaluates whether QSEs set aside a sufficient quantity of un-loaded, ramp-capable, on-line capacity to meet their responsive reserves obligations in real-time.

When interpreting the results of our analysis, it is important to understand the assumptions used in the analysis. To the extent that these assumptions do not consider all relevant factors, it may lead the analysis to over- or under-estimate the QSEs failure to satisfy their obligations:

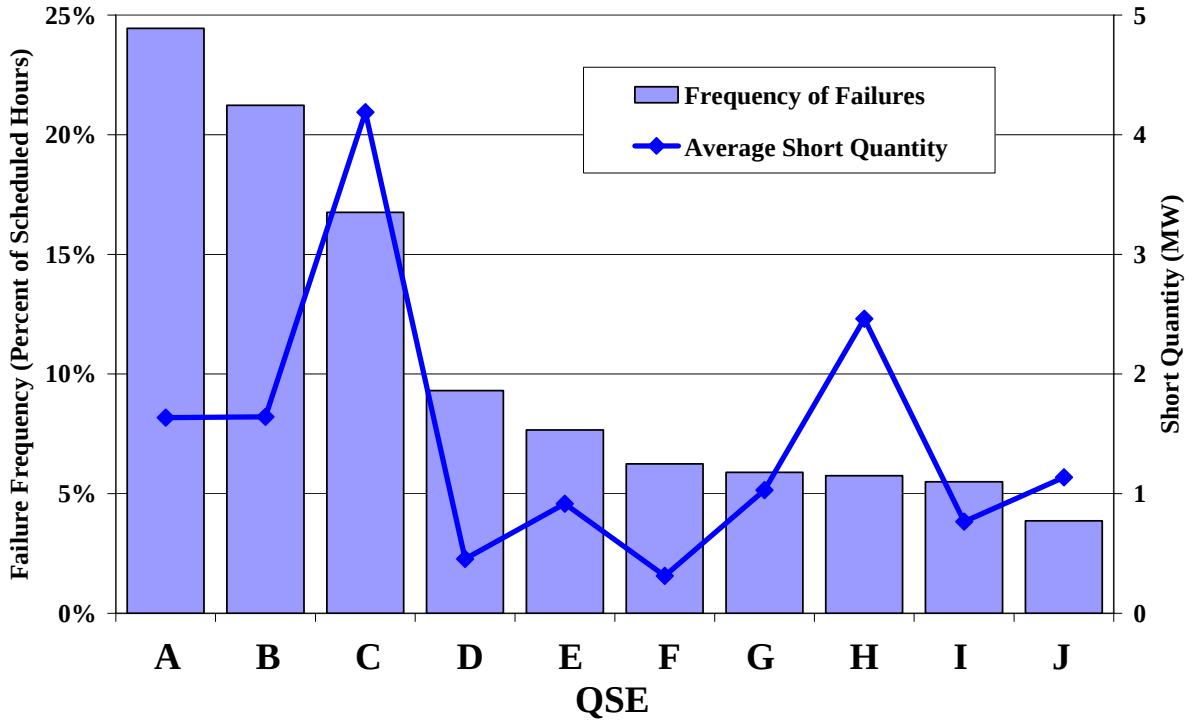
- The headroom of each generator is based on the difference between the maximum output level reported in the resource plan and the actual output of the resource, averaged over the hour.
- The ramp rate used for this analysis is based on the value reported in the resource plan.
- No adjustments are made for units providing OOME service, although QSEs with resources deployed for OOME service may have difficulty satisfying both their energy and ancillary services obligations.
- No adjustments are made for generators that increase output as a result of governor response.

For the purposes of this analysis, we assume that 30 percent of regulation up capability is deployed in each hour requiring QSEs to provide the remaining 70 percent of regulation up from unloaded resources.

Figure 24 shows a summary of the analysis including the ten QSEs that were short of their responsive reserves obligation most frequently during 2005. Shortages were detected by analyzing the ability of each QSE to ramp up their controllable on-line resources within 10 minutes to satisfy their responsive reserve obligation, given the QSE's unit specific ramp rate limitations. The percentage shown on each bar in Figure 24 is the percentage of hours when the

QSE did not satisfy its reserve obligation according to our analysis. The line indicates the average number of megawatts the QSE was short of its obligation averaged across all hours when the QSE was scheduled.[19]

Figure 24: QSEs Failing to Satisfy Their Responsive Reserve Obligations 2005



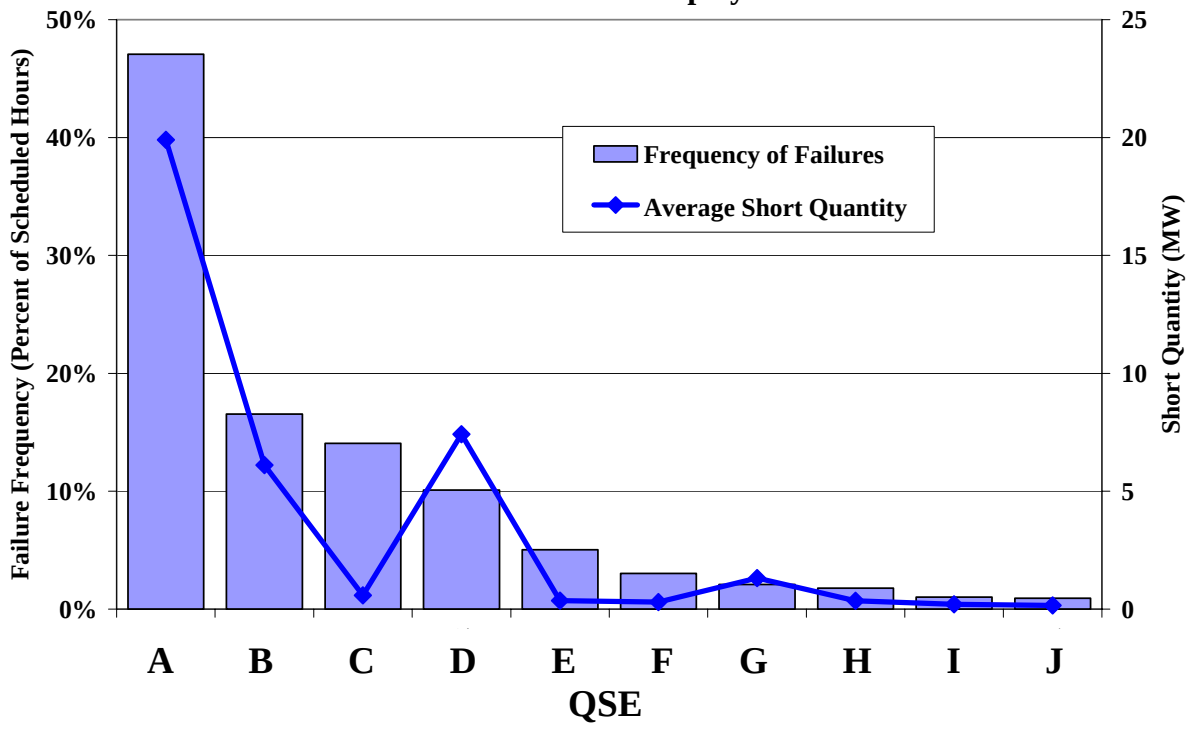
While the ten individual QSEs appear to be short of their reserves obligations a significant portion of the time they are scheduled to provide reserves, the shortages by these QSEs account for an average of approximately 1 percent of the market-wide responsive reserves requirement. Given the potential measurement error in identifying the failures and the small average quantities of the failures, these results raise very limited concerns.

The following analysis evaluates whether QSEs are setting aside sufficient capability to meet their non-spinning reserves obligations. Like the prior analysis, Figure 25 shows a summary of our analysis identifying the ten QSEs that were most frequently failing to satisfy their non-spinning reserves obligation during 2005. Shortages were detected by analyzing the ability of each QSE to ramp up their on-line resources and start available off-line resources within 30

[19] For example, if a QSE was scheduled for reserves in two hours and was short 10 MW in one of the two hours, the figure would show a frequency of 50 percent and an average shortage of 5 MW.

minutes to satisfy their non-spinning reserve obligations. The percentage shown on each bar in the following figure is the percentage of hours when the QSE did not satisfy its reserve obligations. The line indicates the average number of megawatts the QSE was short of its obligation in those hours. The analysis is limited to those hours when reserves were not being deployed because such a deployment could make a QSE appear to be short of its required reserves.

Figure 25: QSEs Failing to Satisfy Their Non-Spinning Reserve Obligations In Hours Without Reserves Deployments – 2005



The assumptions used for this analysis are consistent with the assumptions used for the comparable analysis of responsive reserves. However, this analysis assumes that QSEs meet the responsive reserves obligations first, and their non-spin obligations second. As with the previous analysis, the assumptions used in this analysis could lead to over- or under-estimation of individual QSE's failures to provide reserves. Like the results of the prior analysis, the estimated reserve shortages of the ten QSEs generally account for a small share of the non-spinning reserves requirement. Although the previous two analyses indicate that QSEs fail to provide only a small portion of the total reserves requirement, there are incentives for QSEs to use some of their reserves capacity to provide energy. These incentives are particularly strong when the balancing energy market clearing price is very high.

ERCOT estimates unloaded capacity by QSE, both leading up to real-time and in real-time. However, it may not have sufficient information to accurately monitor non-spin and responsive reserve capability by QSE, which requires information on the response rates of the QSE's units, their output levels, their usable capacity, and their eligibility to provide responsive reserves. To improve the performance of the QSEs in meeting their obligations, we recommend that ERCOT monitor the reserves held in real-time and withhold payments corresponding to the quantity of reserves that were not maintained. The ERCOT Board of Directors has approved PRR 590 which would accomplish this by requiring QSEs to submit the AGC status and ramp rate for each resource in the portfolio through telemetry. The process for implementing this objective has not yet been determined.[20]

3. Ancillary Services Conclusions

Ancillary services prices have risen considerably since 2002, consistent with long-term trends in natural gas and electricity prices. Because ancillary services 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. Providers of responsive reserves and regulation can incur opportunity costs when they reduce the output from economic units to make the capability available to provide these services. In late 2004, there was a large increase in ancillary services prices, coincident with an increase in the frequency of price spikes in the balancing energy market.

Although ancillary services prices have risen over the last few years, the impact has been partly mitigated by reductions in the required quantities of regulation. In 2002, ERCOT required approximately 3,000 MW of combined up and down regulation. By 2005, the requirement was reduced to an average of 1,950 MW during ramping hours and 1,400 MW during non-ramping hours. This has *directly* reduced regulation costs by reducing the overall quantity scheduled, either through bilateral arrangements or through the day-ahead auction. This has also *indirectly* reduced regulation costs by reducing the clearing prices of regulation that would have prevailed under higher demand levels for regulation.

[20] Operating Guide Revision Request ("OGRR ") 165, which was designed to implement PRR 590, has not yet been approved.

In this report, we compare the amounts of capacity scheduled to provide operating reserves to the quantities of capacity that are actually available in real time. In generally, we find that the capacity available to provide reserves in real time far exceeds the quantities scheduled to meet the operating reserves requirements. Although we find that individual QSEs sometimes do not set aside sufficient capacity in real-time to satisfy their operating reserves obligations, the magnitude of these shortages is very small relative to the market's required quantities of operating reserves. Stakeholders are currently in the process of determining how to implement PRR 590, which directs ERCOT to monitor the compliance of individual QSEs with their operating reserves obligations. We also note that QSEs are sometimes short of their responsive reserves obligations due to the provision that does not allow more than 20 percent of the capacity of a particular resource to provide responsive reserves.

In October 2005, ERCOT modified the day-ahead procurement process for ancillary services so the markets for regulation, responsive reserves, and non-spinning reserves clear simultaneously. By running a simultaneous auction for the four services, it is possible to clear the market with the least expensive set of available offers. This change is likely to result in more efficient prices for ancillary services since they will better reflect the opportunity costs of not providing the other services.

ERCOT continued to incur relatively high costs for reserves and regulation compared with other markets in 2005. This occurs even though sufficient excess capacity is usually on-line and available to the balancing market. 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 2005, 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 it results in an allocation of resources to provide ancillary services that is suboptimal. The only market with comparable responsive reserves prices is PJM, which also does not jointly-optimize the procurement of reserves and energy.

C. Net Revenue Analysis

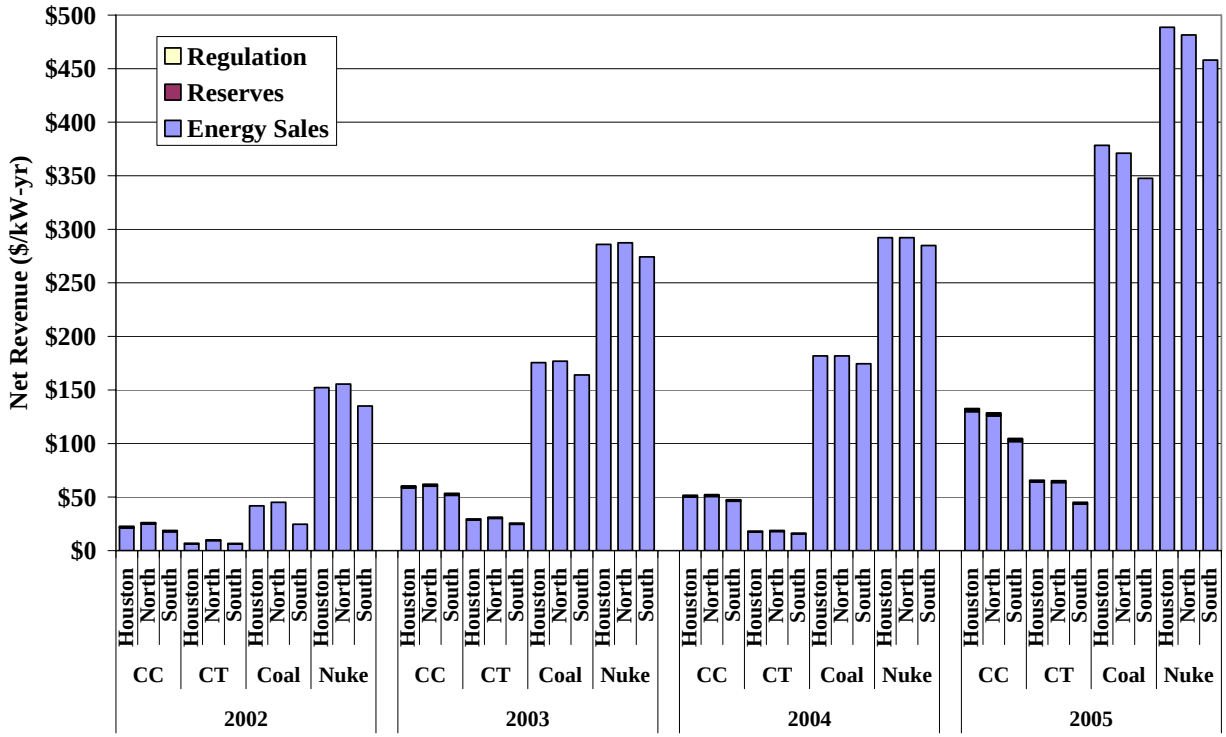
Net revenue is defined as the total revenue that can be earned by a generating unit less its variable production costs. Hence, it is the revenue in excess of short-run operating costs and is available to recover a unit's fixed and capital costs. Net revenues from the energy, operating reserves, and regulation markets together provide the economic signals that govern suppliers' decisions to invest in new generation or retire existing generation. In a 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:

- New capacity is not needed because there is sufficient generation already available;
- Load levels, and thus energy prices, are temporarily below long-run expected levels (this could be due to mild weather or other factors); or
- 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 between 2002 and 2005 by various types of generators in each zone.

Figure 26 shows the results of the net revenue analysis for four types of units. These are: (a) a gas combined-cycle, (b) a combustion turbine, (c) a new coal unit, and (d) a new nuclear unit. In recent years, most new capacity investment has been in natural gas-fired technologies, although high prices for oil and natural gas have caused renewed interest in new investment in coal and nuclear generation. For the gas-fired technologies, 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 that it is available (i.e., when it is not incurring an planned or forced outage). For coal and nuclear technologies, net revenue is calculated by assuming that the unit will produce at full output. The energy net revenues are computed based on the balancing energy price in each hour. Although most suppliers would receive the bulk of their revenues through bilateral contracts, the spot prices produced in the balancing energy market should drive the bilateral energy prices over time.

Figure 26: Estimated Net Revenue 2002 to 2005



For purposes of this analysis, we assume heat rates of 7 MMbtu per MWh for a combined cycle unit, 10.5 MMbtu per MWh for a combustion turbine, and 9 MMbtu per MWh for a new coal unit. We assume variable operating and maintenance costs of $4 per MWh for the gas units and $1 per MWh for the coal unit. We assume variable costs of $5 per MWh for the nuclear unit. For each technology, we assumed a total outage rate (planned and forced) of 10 percent.

The highest net revenues were in the North and Houston zones while lowest net revenue levels were in the South zone. Because the net revenues for the Northeast and West zones fall within the range of the other three zones, we do not show their net revenues in the figure for legibility. Although the analysis indicates that a generator operating in the North zone or in Houston would have earned more net revenue than a generator in the South zone, the relative costs of investment in these zones are important in determining the most attractive locations for new investment.

Some units, generally those 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 Capacity, and Reliability Must Run payments). This source of revenue is not considered in this analysis. The analysis also includes simplifying assumptions

that can lead to over-estimates of the profitability of operating in the wholesale market. The following factors are not explicitly accounted for in the net revenue analysis: (i) start-up costs, which can be significant; and (ii) minimum running times and ramp restriction, which can prevent the natural gas generators from profiting during brief price spikes. Despite these limitations, the net revenue analysis provides a useful summary of signals for investment in the wholesale market.

Figure 26 shows that the estimated net revenue for all technologies grew significantly from 2002 to 2003 and again from 2004 to 2005. The net revenue increases observed in the figure are the result of the higher energy prices in 2005 that were due to: rising natural gas prices and relatively frequent price spikes associated with balancing energy shortages. The higher net revenues in 2005 make it more likely that a new gas-fired generating unit would have earned sufficient net revenue to make the investment profitable. Based on our estimates of investment costs for new units, the net revenue required to satisfy the annual fixed costs (including capital carrying costs) of a new gas turbine unit is approximately $60 to $75 per kW-year.[21] The estimated net revenue for a new gas turbine in 2005 is slightly lower than this level. For a new combined cycle unit, the estimated net revenue requirement is approximately $95 to $125 per kW-year,[22] which is consistent with the estimated net revenue for this type of unit in 2005. The annual revenue requirements above are for new construction. Other types of projects may have substantially lower investment costs, such as projects to upgrade existing facilities or to re-power old sites.

Prior to 2003, net revenues were well below the levels necessary to justify new investment in coal and nuclear generation. However, high natural gas prices have allowed energy prices to remain at levels high enough for these technologies to be economically viable. The production costs of coal and nuclear units did not change significantly over this period, leading to a dramatic rise in net revenues. The annual fixed costs (including capital carrying costs) are estimated at

21 This uses the same assumptions except for EIA estimated overnight costs of $367 per kW to install a new combustion turbine unit, amortization over 15 years, and a three year phase of construction.

22 This uses the same assumptions except for EIA estimated overnight costs of $556 per kW to install a new combustion turbine unit, amortization over 15 years, and a four year phase of construction.

$190 to $245 per kW-year for a new coal unit[23] and $280 to $390 per kW-year for a new nuclear unit.[24] Net revenues were at the lower ends of these ranges in 2003 and 2004 and exceeded them by a considerable margin in 2005. Thus, it is not surprising that some market participants are expressing interest in build new baseload facilities in ERCOT.[25] However, these results should be tempered by the fact that there are likely additional costs for these technologies that are not included in our generic cost estimates, including the costs associated with the nuclear waste disposal.

Given the surplus of capacity that existed in ERCOT in 2005, it is surprising that net revenues in 2005 exceeded the levels needed for new investment. Several factors explain the net revenue levels. First, a substantial share of the capacity in ERCOT is made up of older gas and oil-fired steam turbines that sets prices at levels higher marginal costs of the new units. Although this capacity helps satisfy ERCOT's resource needs, it may be replaced in the future with plants that are either more fuel efficient or that have lower input fuel costs. Second, as we discuss in the next section, the current market design does not facilitate full utilization of the economic resources in ERCOT. Price spikes frequently occur in the balancing market while substantial amounts of on-line capacity remain unutilized for energy or ancillary services. Third, a supplier raised its offer prices well above estimated marginal costs for a significant portion of its portfolio during the summer of 2005 resulting in higher balancing energy prices. These offers are reviewed in the final section of this report.

Although estimated net revenue grew considerably in 2005 to levels that would likely justify new investment, there are other factors that determine incentives for new investment. First, market participants must anticipate how prices will be affected by the new capacity investment, future load growth, and increasing participation in demand response. Second, net revenues can be

---

23 This is based on the Energy Information Agency's ("EIA") estimate of $1,167 per kW overnight cost to install a new coal unit with scrubber technology. We assume 50 percent debt financing at a rate of 8 to 10 percent, a cost of capital between 13 and 15 percent, amortized over 20 years, and four years of construction before the start of operation. We assume that the project is financed evenly across the construction phase and a 38 percent tax rate.

24 This uses the same assumptions except for EIA estimated overnight costs of $1,744 per kW to install a new nuclear unit, amortization over 30 years, and a six year phase of construction.

25 NRG Energy announced plans to add 2,700 MW at the STP nuclear plant and 800 MW at the Limestone coal plant in a June 21, 2006 press release.

inflated when prices clear above competitive levels as a result of market power being exercised. Thus, a market participant may be deterred from investing in new capacity if it believes that prevailing net revenues are largely due to an exercise of market power that would not be sustainable after the entry of the new generation. Third, the nodal market design that ERCOT plans to implement in 2009 will have an effect on the profitably of new resources. In a particular location, nodal prices could be higher or lower than the prices in the current market depending on the pattern of transmission losses and congestion.

To provide additional context for the net revenue results presented in this section, we also compared the net revenue for natural gas-fired technologies in the ERCOT market with net revenue in other centralized wholesale markets. Figure 27 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,[26] and (e) the PJM. The figure includes estimates of net revenue from energy, reserves and regulation, and capacity. ERCOT does not have a capacity market, and thus, does not have any net revenue from capacity sales.[27]

[26] The ISO-New England revised its methodology in 2005 to include estimated revenues from its forward reserves market for the 10,500 BTU/kWh unit. Although this market also existed in 2004, the figures for 2004 do not include forward reserves revenue.

[27] The California ISO does not report capacity and ancillary services net revenue separately, so it is shown as a combined block in Figure 27. 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.

Figure 27: Comparison of Net Revenue between Markets 2002 to 2005

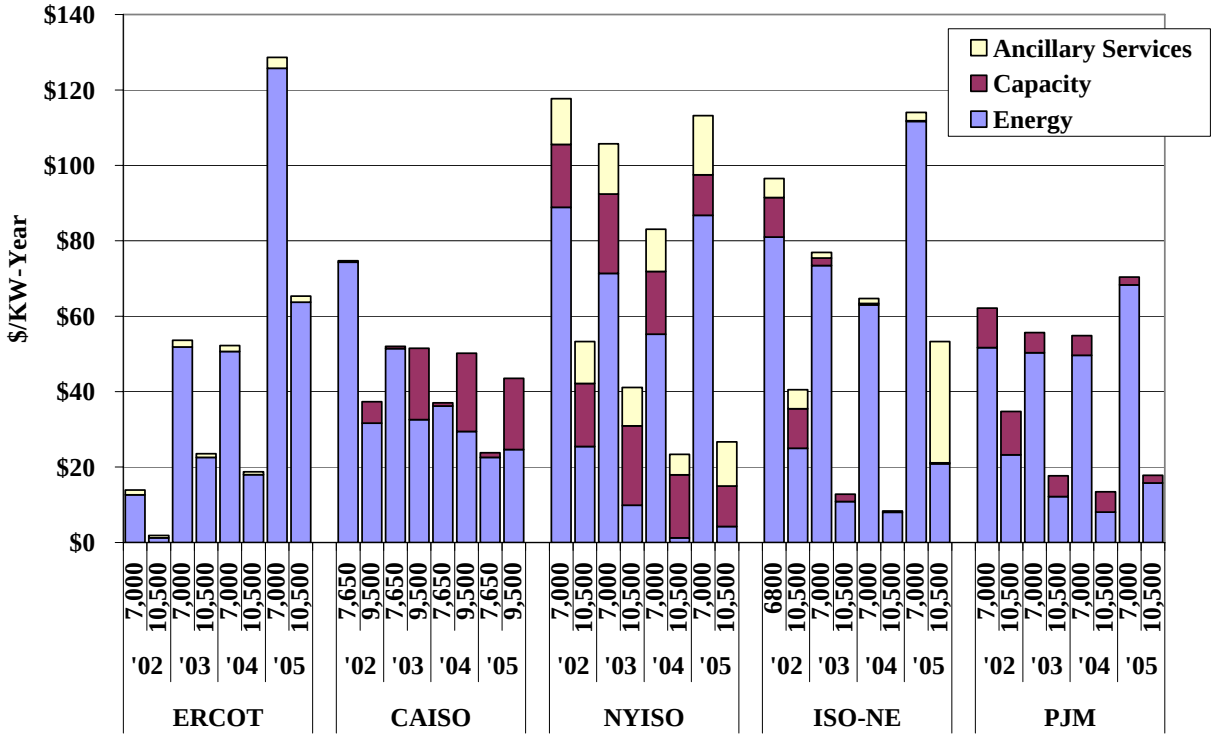


Figure 27 shows that net revenues rose considerably from 2004 to 2005 throughout the U.S., with the exception of California. Most of these areas experienced a mild summer in 2004 and much higher demand in 2005. Of the five markets shown above, ERCOT exhibited the largest increase in net revenues from 2004 to 2005. This difference can be explained by several 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 to fall for the hypothetical new units that can only burn natural gas. Finally, the increased frequency of price spikes in ERCOT, as discussed above, also contributed to the increase in net revenue.

In 2005, ERCOT and the Northeastern markets exhibited net revenue in a range that might be sufficient to motivate investment in new gas-fired capacity, while net revenue in California and PJM likely would not likely be sufficient to support investment in new capacity. However, the costs of new investment can vary significantly by region due to widely varying costs of land,

access to water and fuel, and other regional factors, such as state and local tax and regulatory costs. In the figure above, net revenues are calculated for central locations in each of the five markets. However, there are load pockets within each market where net revenue, and the cost of new investment, may be higher. Thus, even if new investment is not generally profitable in a market, it may be economic in certain areas.

In a market with efficient pricing, spot price signals should indicate when and where new generation investment is needed and when existing generation should be retired. Under the nodal market design, it will be important to ensure that the market sends efficient signals for new investment and retirement. This is primarily accomplished in one of two ways:

- A capacity market; and/or
- Shortage pricing provisions to ensure that prices rise appropriately in the energy and ancillary services markets to reflect the true costs of shortages when resources are insufficient to satisfy both the energy and ancillary services requirements.

A market with one or both of these provisions can maintain adequate levels of supply without allowing market participants to artificially raise prices by withholding resources. Relying on large participants to raise prices by withholding in order to generate efficient long-term economic signals is inferior to relying on shortage pricing provisions because it is unlikely to provide efficient price signals in each location. One of the most significant benefits of co-optimizing energy with ancillary services in real-time is that it allows for efficient shortage pricing. Clearing prices in a co-optimized market reflect the trade-offs between energy and ancillary services when they system is in shortage and the requirements of both markets. Such a market allows prices to rise sharply during periods of true shortage, providing the long-term economic signals that govern new investment and retirement decisions in generation and transmission facilities.

II. SCHEDULING AND BALANCING MARKET OFFERS

In the ERCOT market, QSEs 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 to sixty minutes before the operating hour. QSEs are also required to submit a resource plan that indicates the units that are expected to be on-line and satisfying their scheduled energy obligations. Under ERCOT's relaxed balanced schedules policy, the load schedule is not required to approximate the QSE's projected load. When a QSE's load 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 load schedule is greater than actual load, its generation is over-scheduled and it will sell the residual in the balancing energy market at the balancing energy price.

The QSE schedules and resource plans are the main 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 load scheduling by QSEs. The analysis focuses on the degree to which load 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. Load Scheduling

In this subsection, we evaluate load scheduling patterns by comparing load schedules to actual real-time load. We focus on the load schedules at two points in time. First, we will refer to the

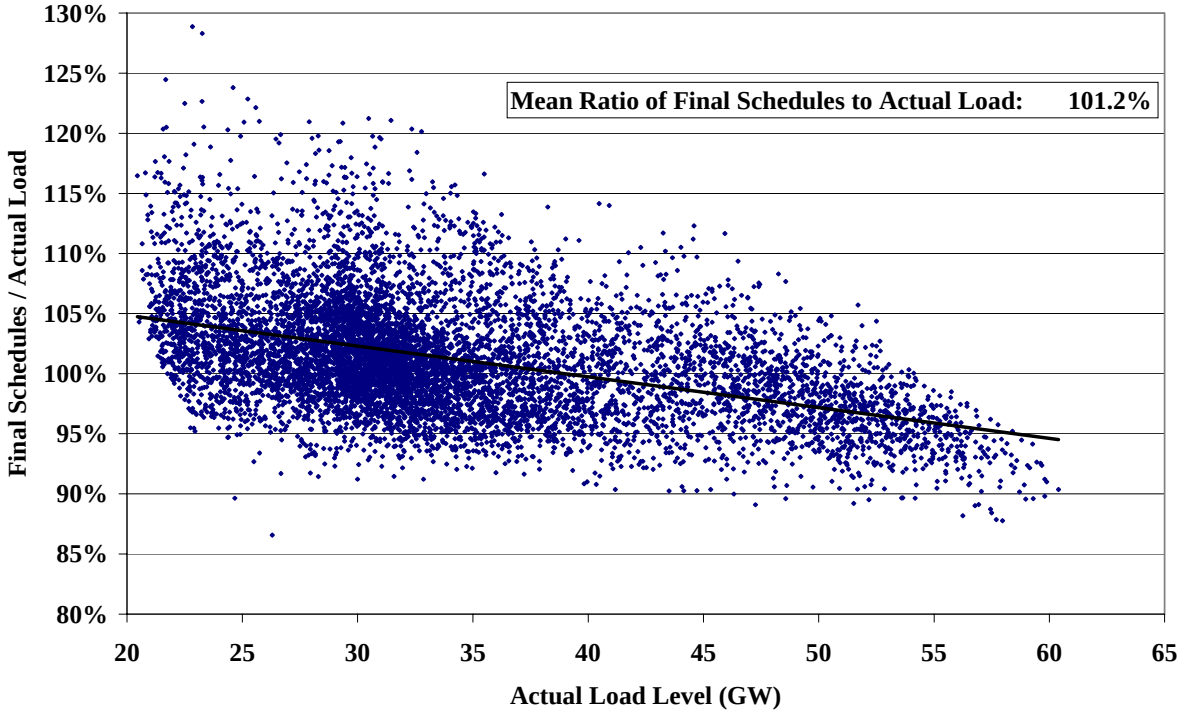
final load schedule, which is the last schedule submitted by the QSE prior to the operating hour. Second, we will refer to the day-ahead schedule, submitted by the QSE in the day ahead.

Relaxed Balanced Scheduling requires QSEs to submit generation schedules to ERCOT that matches scheduled load. It is routine for QSEs to submit quantities with small discrepancies causing load schedules to slightly exceed aggregate generation schedules. This does not result in operational problems because ERCOT schedules balancing energy to meet any unbalanced load in real-time. However, ERCOT encourages QSEs to keep load and generation schedules consistent by charging a small fee for each unbalanced megawatt.

Relaxed Balanced Scheduling does not require that scheduled load be close to actual load. Differences between the two are settled in the balancing energy market. Hence, periods with large levels of under-scheduling creates substantial demand for balancing energy. The analyses in this section of the report evaluate the load scheduling patterns and how those patterns affected energy prices and uplift.

To provide an overview of the scheduling patterns, Figure 28 shows a scatter diagram that plots the ratio of the final load schedules to the actual load level during 2005. The ratio shown in the figure will be greater than 100 percent when the final load schedule is greater than the actual load. Therefore, in general, the figure shows that final load schedules come very close to actual load in the aggregate, as indicated by an average ratio of the final load schedules to actual load of 101.2 percent. However, the figure also includes a trend line indicating that the ratio of final load schedules to actual load tends to decrease as load rises. In particular, the ratio given by the trend line is above 100 percent for loads under 38 GW and declines to 95 percent at 60 GW of load. The overall pattern shown in the figure above is similar to 2004, which exhibited the same downward trend in final load schedules relative to actual load, although the average ratio was only 99.3 percent.

Figure 28: Ratio of Final Load Schedules to Actual Load
All ERCOT – 2005



On average, balancing energy prices are higher and more volatile at high load levels, although the previous subsection showed that spikes can occur under all load conditions. Market participants that are risk averse might be expected to schedule forward to cover their load during high load periods rather than reducing their forward scheduling levels during those periods. There are several explanations for the apparent under-scheduling during high load conditions. 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 greater price risk. Financial contracts or derivatives may be in place to protect market participants from price risk in the balancing energy market, such as a contract for differences. Second, market participants who own generation can offer their expensive generation into the market to cover their load needs if balancing energy market prices are high but otherwise allow their load obligations to be met with lower priced balancing energy.

Figure 29 is a further analysis of final load schedules that shows the ratio of final load schedules to actual load evaluated at five different load levels for each of the ERCOT zones.

Figure 29: Average Ratio of Final Load Schedules to Actual Load by Load Level All Zones – 2005

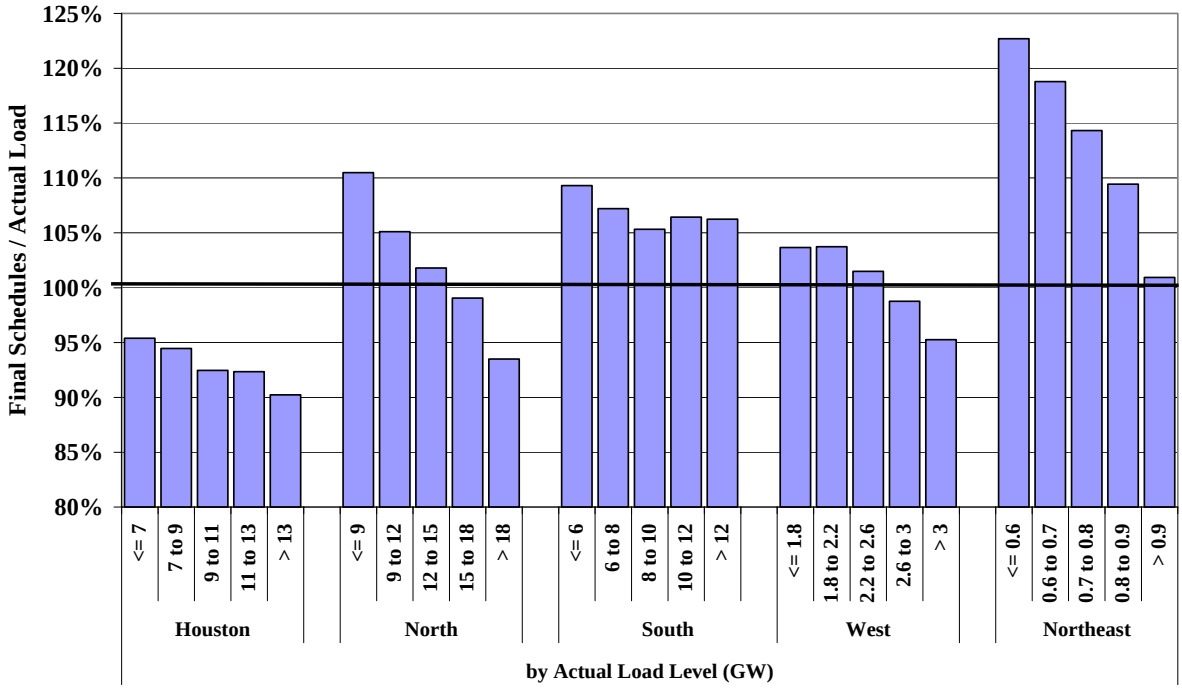


Figure 29 shows that:

- The final schedule quantity decreases in each of the five zones as actual load increases, with the exception of the South zone where a slight increase in the ratio occurs at high load levels.
- The South Zone is generally over-scheduled, although the ratios decline slightly as load increases.
- The Northeast Zone is consistently over-scheduled by a large margin. However, since the Northeast Zone accounts for less than 3 percent of ERCOT load, the total amount over-scheduled is usually less than 300 MW.
- Houston is under-scheduled at each load level, ranging from 5 percent at lower load levels up to 10 percent at high load levels.

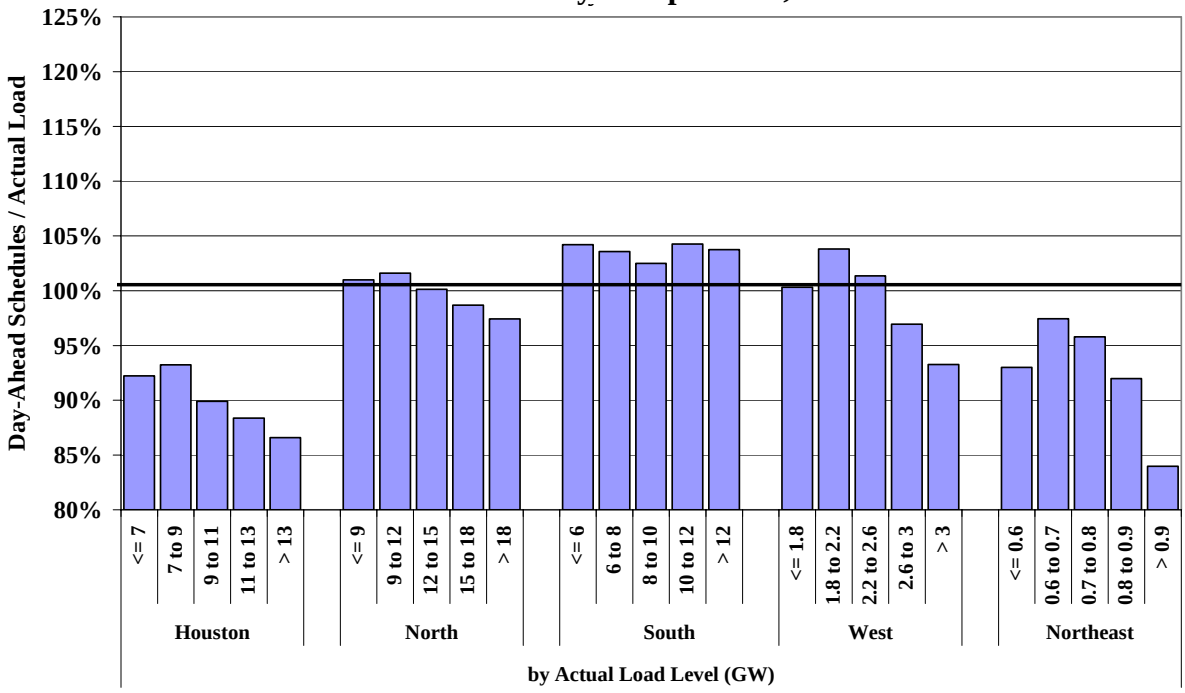
The result of these scheduling patterns is that the QSEs in Houston are net buyers of balancing energy to the extent that they do not offer generation in the balancing energy market to cover their deficits. In contrast, QSEs in the Northeast Zone, and in the South Zone to a lesser degree,

are net sellers of balancing energy. Thus, the net importing zones seem to under-schedule while the net exporting zones over-schedule.

Persistent load imbalances are not necessarily a problem. It can reflect the fact that some suppliers schedule energy from resources they expect to be economic in the balancing energy market when they have not already sold the power in a bilateral contract. Rather than selling power to the balancing energy market through deployments in the balancing energy market, they sell through load imbalances. This poses no operational concerns and is a mechanism by which some suppliers may more fully utilize their portfolio.

We next evaluate the day-ahead load schedules relative to actual load in Figure 30. The figure is analogous to Figure 29. It shows the ratio of day-ahead load schedules to actual load by load level for each of the five zones in ERCOT. The load levels are divided into five roughly equal groups.

Figure 30: Average Ratio of Day-Ahead Load Schedules to Actual Load by Load Level All Zones – January to September, 2005[28]

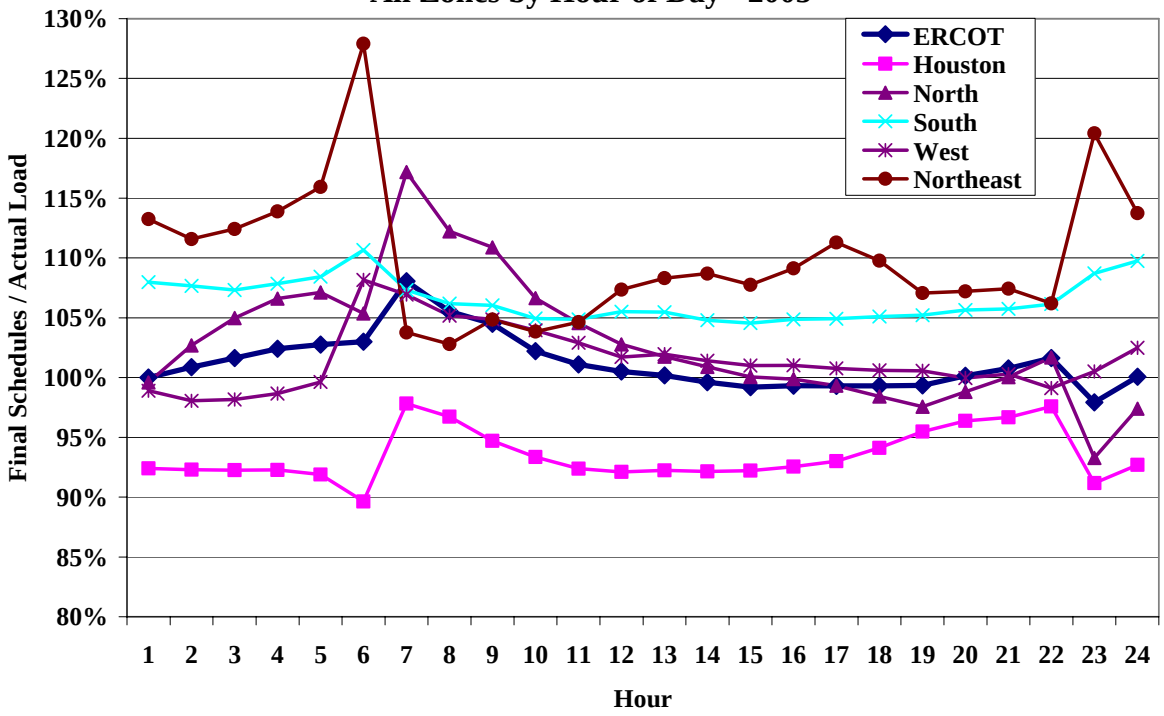


28 Data was not available for the last three months of 2005.

Figure 30 shows that day-ahead scheduling results are generally comparable, both in magnitude and pattern, to the scheduling levels shown in Figure 29 for final load schedules. Day-ahead load schedules in the Houston, North, and West Zones are negatively correlated with actual load levels, but to a lesser degree than the final load schedules. There is also much less scheduling in the Northeast Zone in the day-ahead than in real time. Although there is no obvious explanation for the differences in scheduling patterns between the day-ahead and real-time in the North Zone and Northeast Zone, a more detailed analysis of out-of-merit commitment and dispatch described below provides some insight.

To further analyze load scheduling, Figure 31 shows the ratio of final load schedules to actual load by hour-of-day for each of the five zones in ERCOT as well as for ERCOT as a whole.

Figure 31: Average Ratio of Final Load Schedules to Actual Load All Zones by Hour of Day - 2005



This figure shows that on an ERCOT-wide basis, final schedules are close to actual load (between 100 percent and 103 percent) during hours ending 1 to 6. At hour ending 7, the ratio rises to 108 percent, the highest of any hour. By hour ending 10 through the remainder of the day, the ratio declines to a range between 98 percent and 102 percent.

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 31 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, market participants bear additional price risk in ramping hours (as shown in the prior section), explaining their propensity to schedule a larger portion of their needs during these periods.

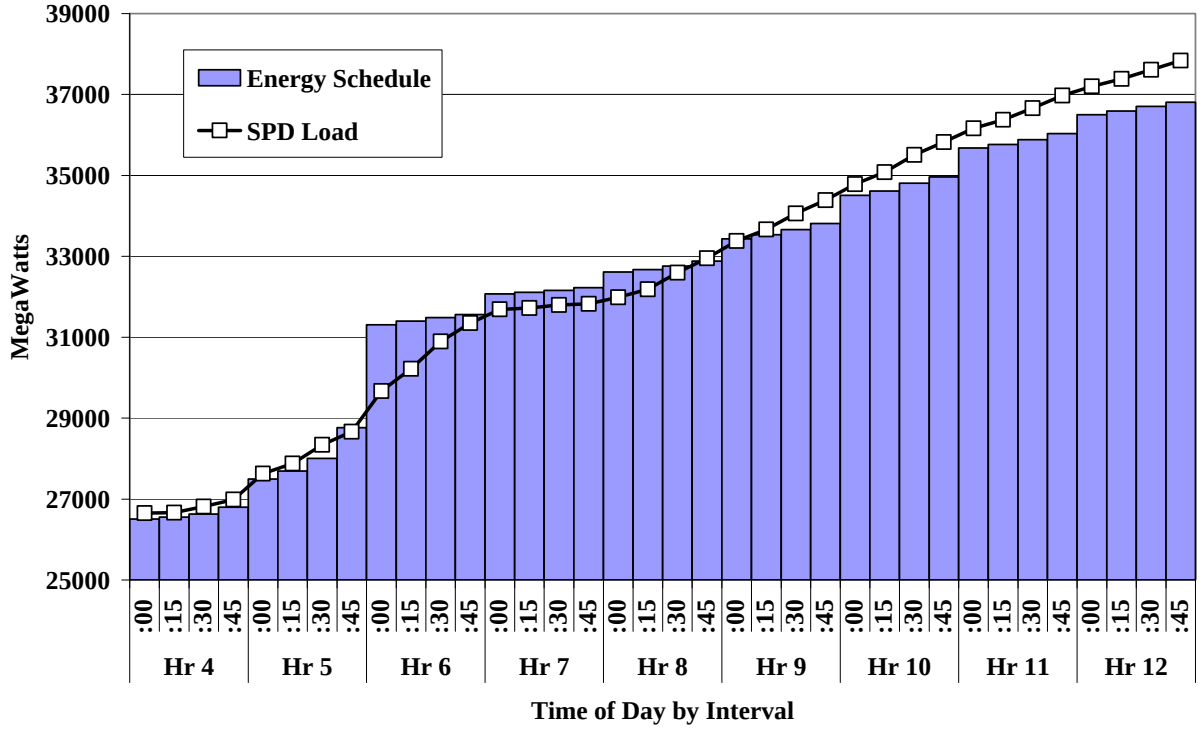
B. Balancing Energy Market Scheduling

In the previous section, we analyzed balancing energy prices and load and found that while balancing energy prices are correlated to real-time load levels, other factors also have substantial effects on balancing energy levels. 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 32 shows average energy schedules and actual load for each interval from 4 AM to 1 PM during 2005.

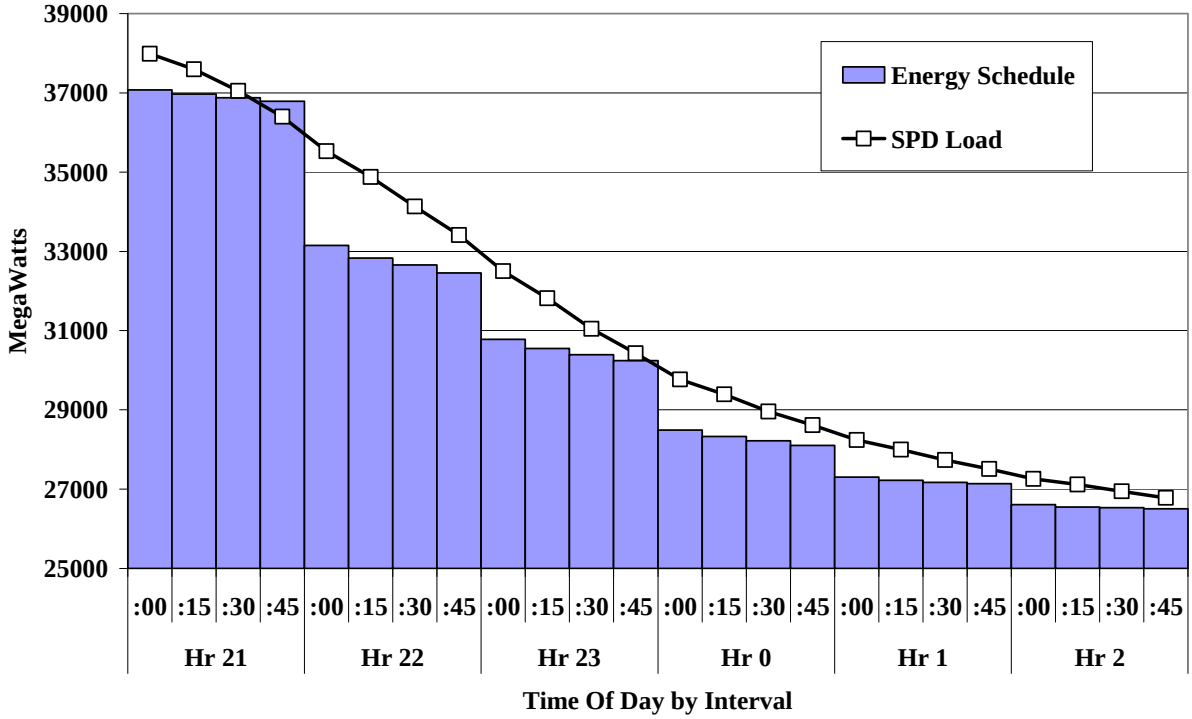
In general, energy schedules that are less than the actual load result in balancing energy purchases while energy schedules higher than actual load result in balancing energy sales. On average, load increases from approximately 27 GW to almost 38 GW in the nine hours shown in Figure 32. The average increase per 15-minute interval is approximately 330 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. This "hump" in the 6 AM to 8 AM timeframe is due, primarily, to the fact that the daily peak occurs in the morning during certain times of year. However, a small hump persists around 6 AM throughout the year.

Figure 32: Final Energy Schedules during Ramping-Up Hours 2005



The increase in load during ramping-up hours is steady relative to the increase in 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 by nearly 3 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 33 shows average energy schedules and load for each interval from 9 PM to 3 AM during 2005.

Figure 33: Final Energy Schedules during Ramping-Down Hours 2005



On average, load drops from approximately 38 GW to less than 27 GW in the six hours shown in Figure 33. The average decrease per 15-minute interval is approximately 480 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 nearly 4 GW from the last interval before 10 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 approximately one-half of the generation in ERCOT is scheduled by QSEs that submit energy schedules that change hourly, while the other half is scheduled by QSEs that submit energy schedules that change every 15 minutes. Deviations between the energy schedules and load scheduled by SPD will result in purchases or sales in the balancing energy market. Specifically, net balancing up energy equals SPD load minus scheduled energy. Hence, Figure 32 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 33 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 32 and Figure 33). This analysis is similar to that shown in Figure 13 and Figure 14, except instead of showing balancing energy prices relative to load, we show balancing energy prices relative to balancing energy deployments. Figure 34 shows the analysis for the ramping-up hours.

Figure 34: Balancing Energy Prices and Volumes
Ramping-Up Hours – 2005

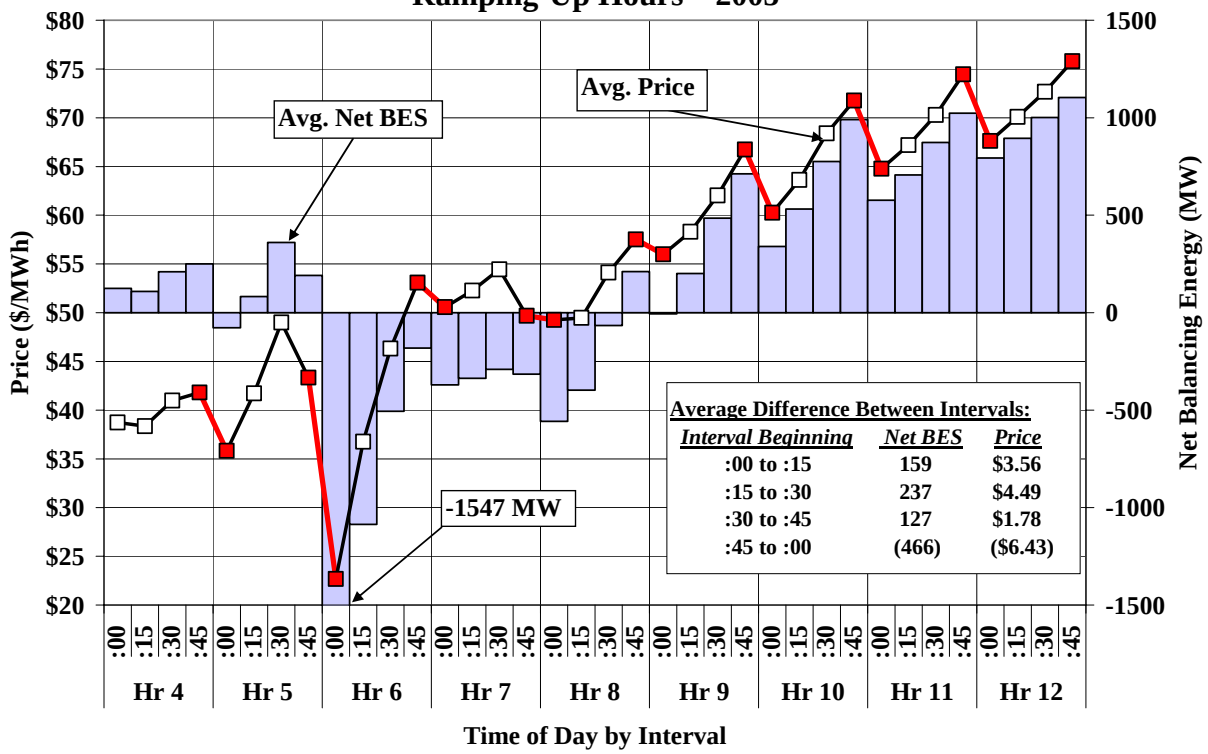
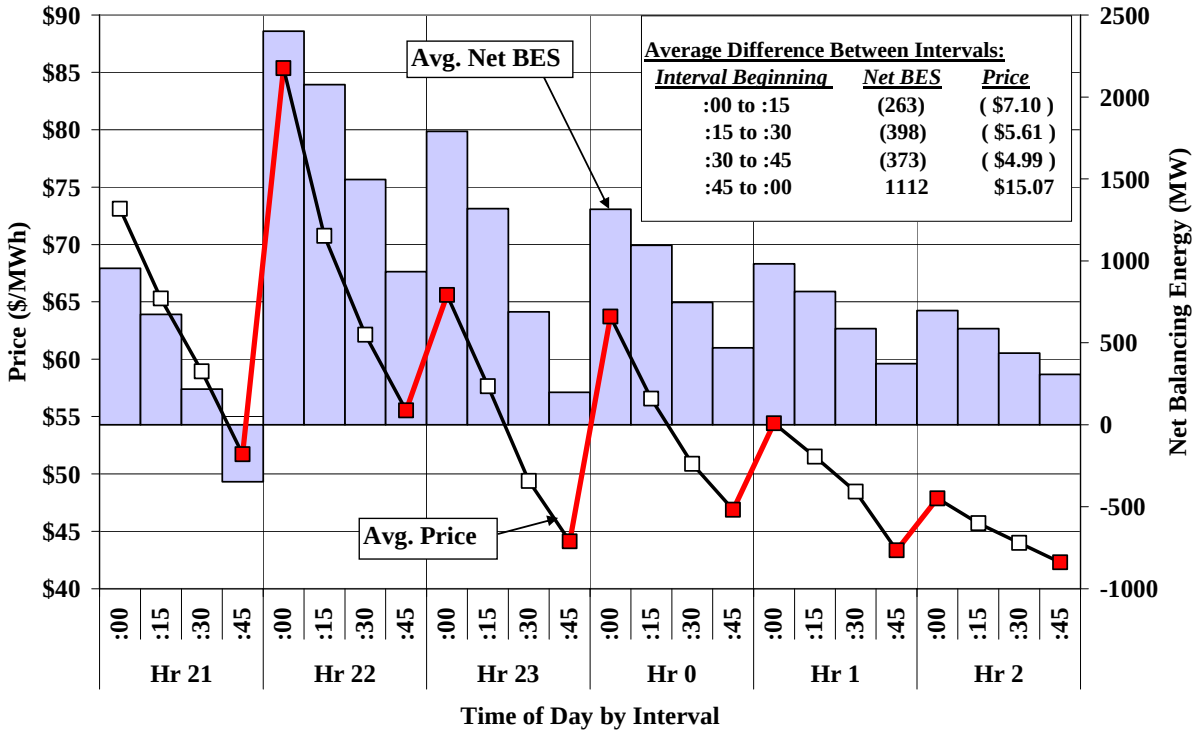


Figure 34 reveals two key aspects of the balancing energy market. First, as discussed above, balancing energy prices are highly correlated with balancing energy deployments. Second, there is a distinct pattern of increasing purchases during the hour. At the beginning of the hour, purchases tend to be 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 35 shows the same analysis for the ramping-down hours.

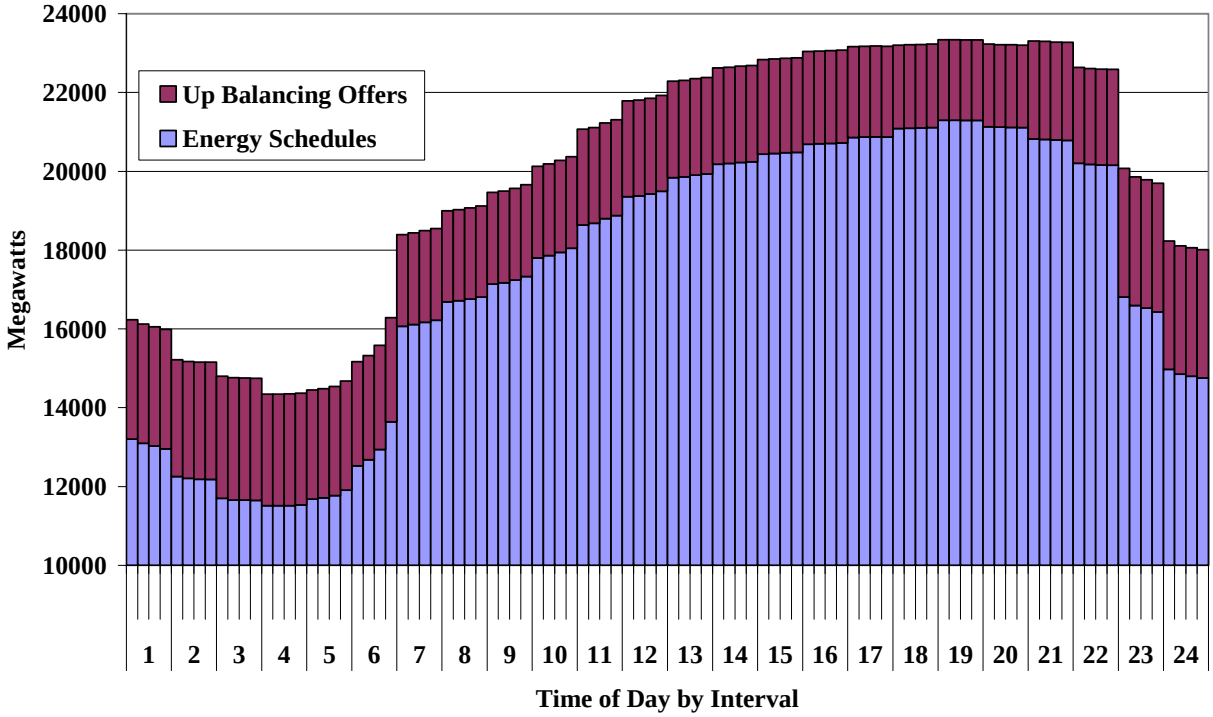
Figure 35: Balancing Energy Prices and Volumes
Ramping-Down Hours – 2005



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. We reviewed QSE scheduling data from 2005 and found that the two largest suppliers in ERCOT tend to schedule much more flexibly than the aggregate schedules of other QSEs. The following two figures analyze the scheduling patterns of the two largest QSEs compared to all other QSEs by interval over the entire day. Figure 36 shows the average quantity of energy schedules and balancing-up offers in 2005 for all QSEs in ERCOT except the largest two.

Figure 36: Final Energy Schedules and Balancing-up Offers
All QSEs except the Largest Two, 2005



This figure shows that energy schedules change modestly on a 15-minute basis, but in relatively large steps from hour to hour. A close examination of the energy scheduling data indicated that five of the QSEs and sub-QSEs in this group scheduled flexibly on a 15-minute basis, at least during ramping hours. However, these five account for approximately 20 percent of the energy schedules shown in the figure above, leaving the majority of QSEs scheduling only on an hourly basis. It is primarily the scheduling patterns by QSEs that schedule on an hourly basis that result in the balancing energy deployments and prices shown in Figure 34 and Figure 35.

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 approximately 2,425 MW on average from the last interval of hour ending 6 to the first interval of hour ending 7, an increase of 18 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 3,340 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 1,460 MW and 1,550 MW, respectively. These large hourly changes in energy schedules are a primary determinant of the balancing energy price fluctuations shown in this section.

Figure 37 shows the energy scheduling patterns of the two largest QSEs, which account for approximately one half of all scheduled energy.

Figure 37: Final Energy Schedules and Balancing up Offers
Largest Two QSEs, 2005

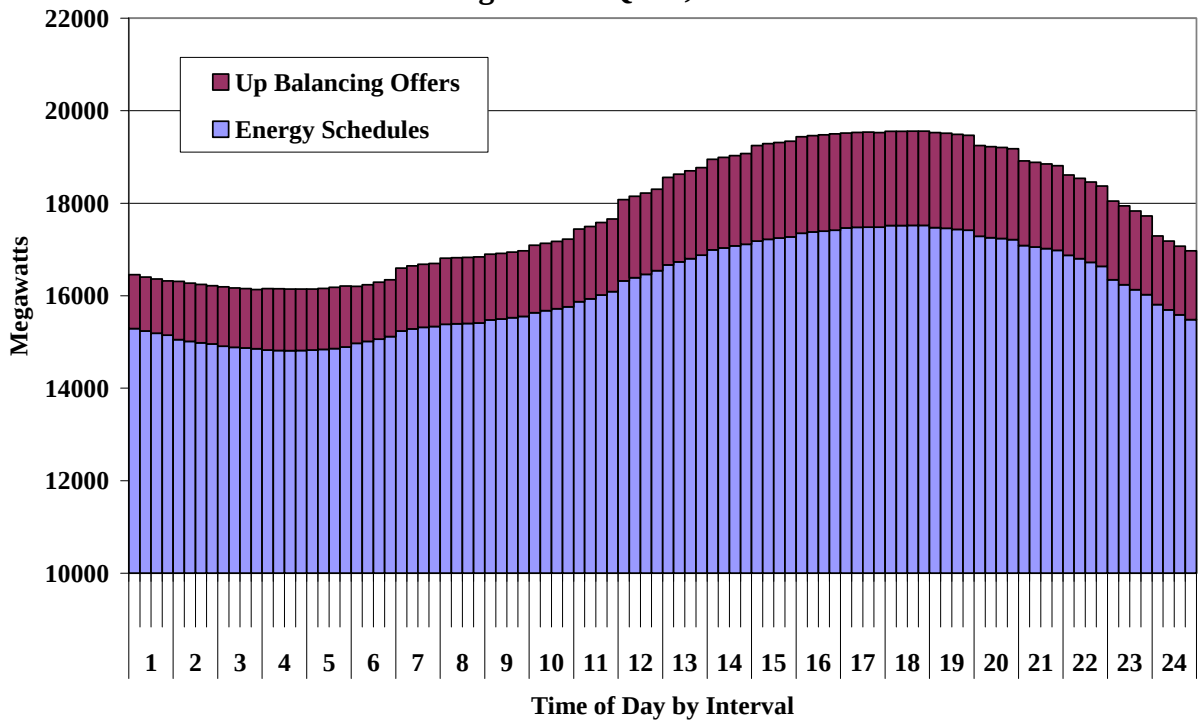


Figure 37 shows that the largest QSEs tend to schedule much more flexibly than the other QSEs. These two QSEs fully utilize the capability to schedule energy on a 15-minute basis. Like the schedules of the other QSEs, these schedules show a shift at the beginning of the peak bilateral contracting period from hour ending 7 to hour ending 22. However, in contrast to other QSEs, the large QSE's energy schedules are relatively flat in the first interval of hour ending 7. While other QSEs decrease their schedules significantly after hour ending 22, these large QSEs show a relatively smooth progression from hour ending 22 to 23.

The large QSEs are in a position to take advantage of profitable arbitrage opportunities that occur at the beginning and end of the peak bilateral contracting period. It allows them to balance

up just before and after the peak bilateral contracting period from hour ending 7 to hour ending 22 and balance down just after it starts and before it ends. Although the large QSEs do not fully 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 improve the performance of the balancing energy market.

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 energy 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.

This issue has been cited in three previous reports,[29] and has continued to be a factor in 2005. To address this issue, we have recommended 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. Furthermore, it would facilitate QSEs offering more of their on-line capacity to the balancing market, which would help address the problem that large amounts of on-line capacity are not offered to the balancing market. Protocol Revision Request ("PRR") 600 was written to address this recommendation but the provision of the PRR that addresses this recommendation has not been retained by the protocol revision subcommittee. This issue should not continue to be a problem under the nodal market design since resource specific offers will not be interpreted as a deviation from an energy schedule.

[29] See 2003 SOM Report, Assessment of Operations, and 2004 SOM Report

C. Portfolio Ramp Limitations

The volatility of the balancing energy 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 energy 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.

In many cases, the lack of ramp capable resources offered to the balancing energy market results in unnecessary price spikes (as well as large negative prices). This sub-section analyzes three aspects of the current market design that inhibit QSEs from fully utilizing the ramp capability of their portfolio. These are: (1) portfolio ramp rates; (2) portfolio level rather than unit level dispatch; and (3) lack of coordination between energy schedules and ramping.

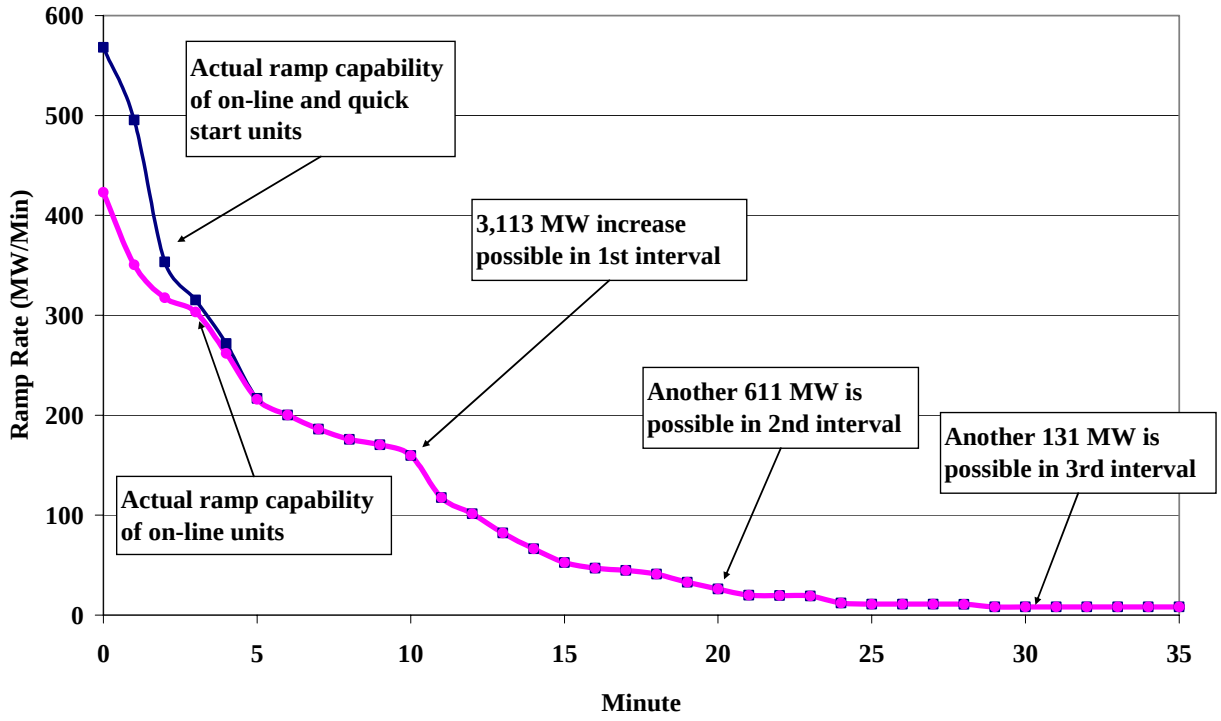
1. Portfolio Ramp Rates

Frequently, market participants are prevented from fully deploying their portfolio by the physical ramp limitations of the resources in their portfolio. To manage this physical limitation, QSEs submit offers with portfolio ramp rates that represent the physical ramp capability of their units on a portfolio basis. Market participants with competitive incentives should submit the highest possible ramp rates that are physically feasible while respecting their ancillary services obligations. Thus, a market participant must estimate the ability of its resources to increase output, which varies as the level of output increases. The following is an analysis that estimates the change in ramp capability for available capacity in ERCOT as resources are deployed upward.

Figure 38 shows two estimates of the physical ramp capability for the dispatchable resources in ERCOT for a sample hour during 2005: (i) including the capability of on-line and quick start resources, and (ii) including the capability of only on-line resources. For this analysis, each unit

is assumed to start out at the planned generation level in the real-time resource plan and increase its output according to the unit's ramp rate specified in the real-time resource plan until the unit reaches its high sustainable limit ("HSL"). Capacity is set aside for each QSE to satisfy its regulation up and responsive reserve obligations.[30]

Figure 38: Physical Ramp Capability of On-Line and Quick Start Resources
March 15, 2005 – Midnight



The figure shows that the ramp capability for all on-line and quick-start resources in ERCOT is 568 MW per minute before any ramping takes place. After one minute of ramping, the ramp capability drops to 496 MW per minute. And after ten minutes of ramping, the ramp capability drops to 160 MW per minute. When an individual resource in ERCOT reaches its HSL, the resource is no longer capable of additional ramp, which implies that the ramp capability of the ERCOT market diminishes as the market increases output. For this reason, both lines in the figure are decreasing over time.

30 This is done by allocating the obligation to its regulation capable units in proportion to ramp capability and headroom. In addition, a QSE capable of ramping 100 MW in a ten-minute period that has a 40 MW responsive reserves obligation can only ramp 60 MW in a ten-minute period and still provide reserves. Thus, 40 MW of capacity capable of ramping in 10 minutes must be set aside by the QSE. The analysis allocates responsive reserves to the most expensive units in the portfolio, wherever possible. The generic resource costs described in Section 5 of the ERCOT Protocols are used to rank generators in order of cost.

In each market interval, QSEs are supposed to move to their instructed dispatch level over a ten-minute period. Thus, the capacity that is physically deployable in one interval is the sum of what can be ramped in the first ten minutes (as shown in the figure above). This starts at 568 MW per minute but decreases to 160 MW per minute after ten minutes for a total of 3,113 MW in the first interval. However, only 611 MW can be deployed in the second interval, and just 131 MW is deployable in the third interval.

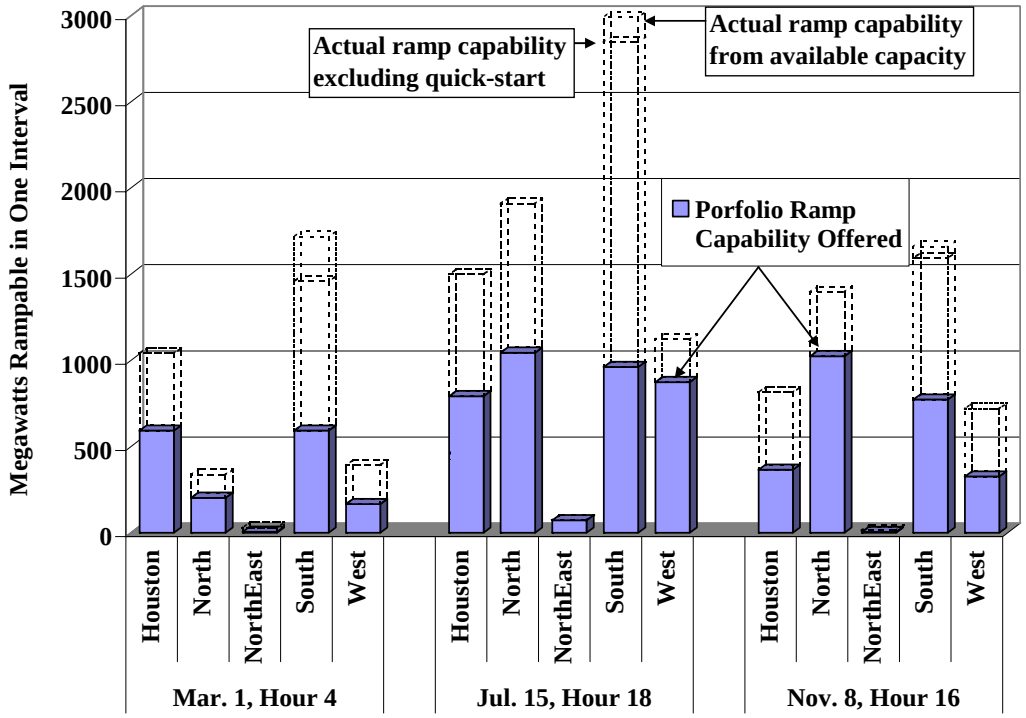
To limit changes in dispatch from one interval to the next, market participants submit a ramp rate for their portfolio in a particular zone. Ideally, market participants could submit a portfolio ramp rate that decreases as production increases to reflect the physical reality shown in Figure 38. However, market participants offer a single constant ramp rate for all capacity above the energy schedule. If they set their ramp rate at a level that allows the maximum feasible deployment in the first interval, it will not be feasible for most QSEs to ramp at that rate in the second interval.

To avoid being over-deployed because of the constant portfolio ramp rate, QSEs have at least three options. First, QSEs can simply lower their portfolio ramp rate to a level that is achievable over all four intervals of the hour. This may allow all capacity to be deployed in one hour, but significantly diminishes the efficient response of the balancing energy market to transient price spikes. Second, QSEs can lower their offer quantities to the amount of capacity that can be physically deployed in two intervals. This allows them to offer a higher ramp rate, but leads to QSEs not offering all of their capacity. Third, QSEs can make multiple portfolio offers, each representing only a fraction of their portfolio. This does not solve the problem, but reduces its magnitude. To address the underlying issue, ERCOT could modify its market software to allow QSEs to offer multiple portfolio ramp rates covering different portions of their offer so that they have the flexibility to offer a larger portion of their available energy into the balancing energy market.

It is not possible to measure how requiring QSEs to offer a constant ramp rate will impact their final offers and/or market outcomes. But it will inevitably lead to smaller offer quantities and lower portfolio ramp rate offers. 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 39, we compare the portfolio ramp rates to actual physical ramp capability at three points in time during

2005. These three hours were selected because they are representative of a variety of conditions during 2005. They occur in the winter, summer, and fall. One is an off-peak hour at night while the other two occur in the afternoon.

Figure 39: Portfolio Ramp Rates versus Ramp Capability Examples from 2005



The bars in this figure 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 quantifies the effects of the current portfolio offer structure. The effects are greatest in the South zone, where more than half of the physically rampable capacity was not

available based on portfolio offers and ramp rates on all three days. In the remaining zones with significant unloaded capacity, rampable portfolio offers were still significantly lower than the total amount of physically rampable capacity on all three days. While we cannot measure how much of the difference between actual rampable capacity and rampable offers is caused by the lack of flexibility QSEs have in submitting portfolio ramp rates, it is likely that a portion of the difference in Figure 39 is attributable to this factor. Therefore, we recommend that ERCOT consider the costs and feasibility of allowing QSEs to offer multiple ramp rates that vary by output level. Protocol Revision Request 675 has been drafted to accomplish this under the current market design, and it is being considered by the Wholesale Market Subcommittee.

2. Unit-Level Dispatch Would Optimize Ramp Capability

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) which 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.

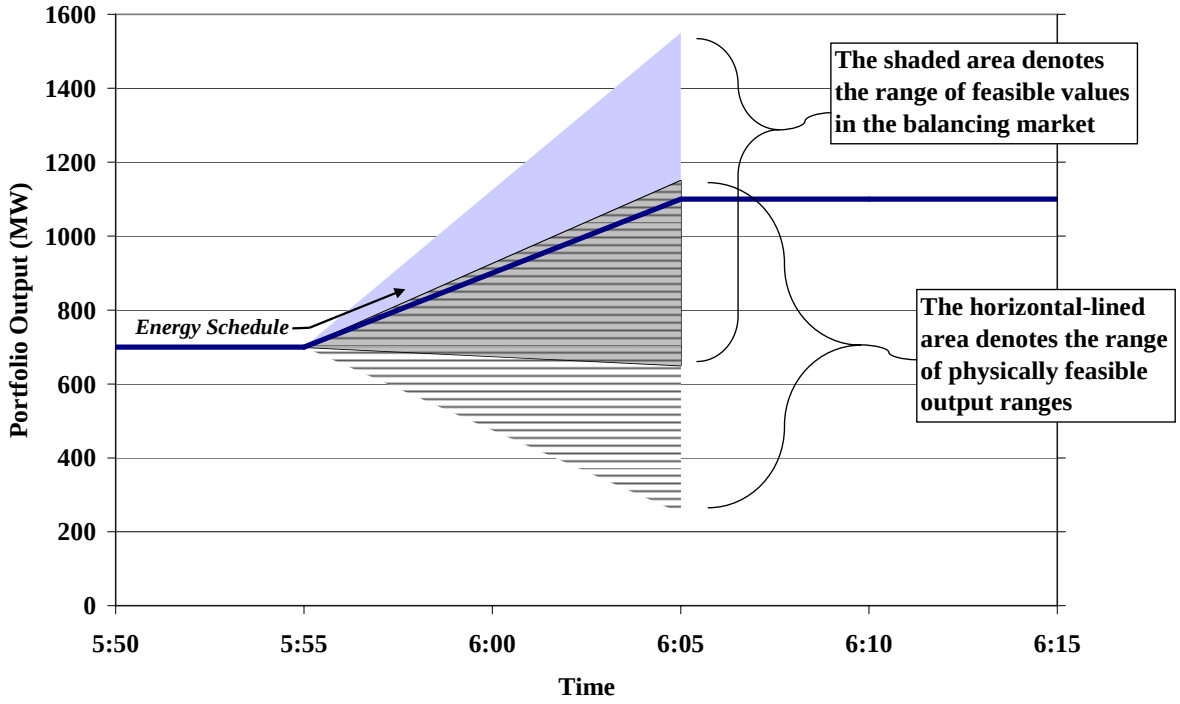
One of the significant benefits of the nodal market design that is currently under development is that it will allow for unit specific bidding and deployment. Allowing suppliers to make offers from specific units with unique ramp rates will provide increased flexibility and more efficient dispatch of the system. Under the current market design, QSEs are able 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. Thus far, several QSEs have chosen to group coal and wind units under separate sub-QSEs, however, most QSEs have not made use of this flexibility.

3. Ramp Constraints in SPD Ignore Schedule Changes

An additional factor that limits the ability of QSEs to submit portfolio ramp rates that make maximum use of their physical resources is that energy schedule changes are not currently considered by SPD when it clears the balancing energy market. QSEs offer balancing-up and balancing-down energy quantities that are constrained by portfolio ramp rate limits relative to balancing energy deployments only in the prior interval.

For instance, if a QSE had 0 MW of cleared balancing-up energy in the first interval and 800 MW of balancing up offers in the second interval with a ramp rate of 45 MW per minute, the QSE could sell up to a total of 450 MW (= 0 MW + 45 MW per minute * 10 minutes) in the second interval. On the other hand, if the QSE's energy schedule increased by 400 MW in the second interval, only 50 MW of balancing energy could physically be deployed in that interval due to the 45 MW per minute ramping limitation. However, the impact of the energy schedule change on physical ramp capability is not recognized by SPD when it deploys balancing energy. The following diagram shows the range of output that SPD *treats as deployable*, indicated by the shaded region, versus the range of output that is *physically deployable*, indicated by the horizontal-lined region.

Figure 40: Treatment of Ramp Rates in the Balancing Market
Comparison with Physical Ramp Capability



The line in the figure above shows that the 400 MW energy schedule change from 700 MW in the first interval to 1100 MW in the second interval would use nearly all of the QSE's physical ramp capability since it would require the output to move up at 40 MW per minute for the ten minutes from 5:55 to 6:05. The physical ramp capability from 5:55 to 6:05 is outlined by the horizontal-lined triangular region, while SPD would treat the QSE as able to move anywhere within the shaded triangular region. Thus, the QSE could ramp at a maximum rate of 45 MW per minute to a total output of 1150 MW at 6:05. However, SPD would treat the QSE as able to move at a rate of 85 MW per minute to a total output of 1550 MW.

Consequently, the QSE is likely to respond by adjusting its ramp rate downward. For instance, because 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). Alternatively, the supplier in this example may simply not offer more than 50 MW of additional balancing energy to the market.

This response by a QSE to lower its portfolio ramp rate offer can prevent suppliers 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, the reduced ramp rate offer would reduce the ability of SPD to send balancing energy deployments that counterbalance schedule changes. Suppose the QSE in the figure above submitted a portfolio ramp rate of 5 MW per minute instead of 45 MW per minute. SPD would constrain the total change in balancing energy deployments to be less than or equal to 50 MW in either the up or down direction, even though the QSE's energy schedule increased 400 MW in one interval. In this case, the SPD could not instruct the supplier to continue to produce the same output as in the prior interval (i.e., the supplier would have to increase output by at least 350 MW) even if the balancing energy prices were well below the production costs of the suppliers' resources.

Second, the portfolio offer curves and ramp rates are set every hour and cannot change by interval. In the example above, the QSE would be limited to a total change in balancing energy deployments of 50 MW in each interval resulting in a maximum increase in balancing-up energy over the hour of 200 MW. In reality, once the QSE increases its output to satisfy its 400 MW schedule change, it is then physically capability of accepting balancing energy deployments of as much as 450 MW in each interval (as long as the QSE has sufficient excess capacity).

To avoid these issues, the QSE could choose to submit a portfolio ramp rate that ignores its changes in energy schedule. However, it would then risk uninstructed deviation penalties and substantial schedule control error 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.

Consequently, many QSEs submit offers to the balancing energy market that are overly restrictive, because either the ramp rate of the offer is too low or the offer quantity is too small. The negative effects of this are seen in Figure 34 and Figure 35, which show that balancing energy prices typically move in a saw tooth pattern during ramping hours. In the morning hours, it is not uncommon for the clearing price in the first interval of the hour to go negative as energy schedules increase much faster than load and overly restrictive offers prevent the balancing model from deploying down balancing energy efficiently. Likewise, in the evening, the clearing

price can spike in the first interval of the hour as energy schedules decrease too rapidly.

To address this issue, we recommend that ERCOT modify its SPD software for the balancing energy market model to account for the ramp capability that is utilized (or created) when the energy schedule increases or decreases. The recognition by SPD that a rise in the QSE's energy schedule increases its capability to balance down (and vica versa) may sometimes provide SPD valuable additional flexibility in making balancing energy deployments that would substantially improve the performance of the balancing energy market.

This recommendation was made in previous reports. This is not likely to be a problem in the nodal market design that ERCOT plans to implement in 2009.[31] In a nodal market, individual units receive dispatch instructions and it should be straightforward to design a ramp constraint that limits changes in production rather than changes in balancing energy sales. Indeed, this is consistent with how ramp constraints are currently used in all of the ISO-operated wholesale markets in the U.S.

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 41, 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 physically on-line or it is from an off-line quick start unit, and if it is not scheduled to provide energy, reserves, or up-regulation. This includes the portion of available capacity under the High Sustainable Limit ("HSL"), and therefore does not include emergency ranges.[32] Unused capacity on renewable resources such as wind turbines are excluded from this category. This data is shown for the peak hour of the day on a monthly average basis from 2003 to 2005 in Figure 41.

31 See 2003 SOM Report, Assessment of Operations, and 2004 SOM Report

32 Although the HSL does not include the emergency range, it may include less flexible operating ranges. For instance, many units are less flexible in their duct firing range. This underscores the benefit of allowing market participants to offer into the balancing market with ramp rates that vary by output level.

Figure 41: Balancing Energy Offers versus Available Capacity[33]
Daily Peak Load Hours – 2003 to 2005

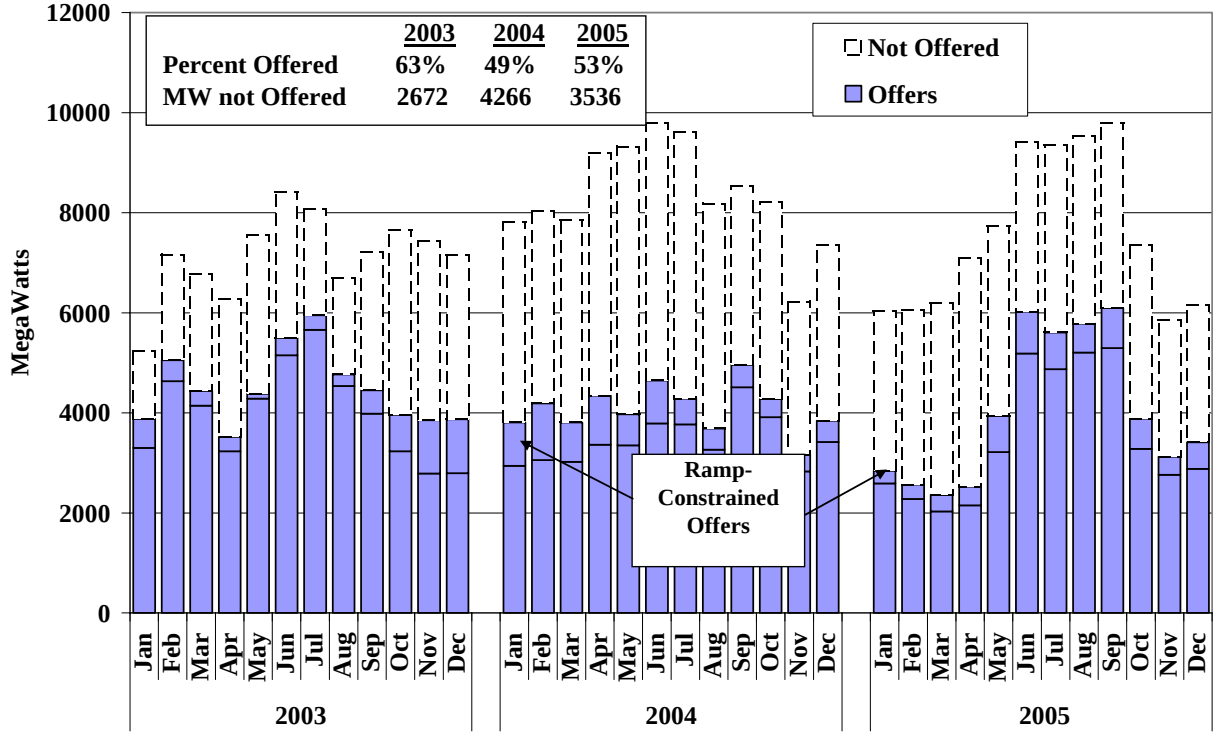


Figure 41 shows the trends over time in quantities of energy available and offered to the balancing energy market. Up balancing offers are divided into the portion that is capable of being deployed in one interval and the portion which would take longer due to portfolio ramp rate offered by the QSE (i.e "Ramp-Constrained Offers). The figure indicates that the amount of available capacity has fluctuated over the past three years. However, the portion that is offered to the balancing energy market has decreased on an annual average basis since 2003. Figure 41 also indicates that the average share of the offer quantity that can be ramped up in one interval (i.e. the portion of the offer below "Ramp-Constrained Offers") is a relatively small portion of the total offer. The figure shows that the portfolio ramp rates offered by market participants have not greatly reduced the portion of offers that are deployable in one interval. This supports the hypothesis discussed in the previous sub-section that one reason participants may not offer all

33 The methodology for this figure differs from a similar chart in the 2004 SOM Report. This analysis assumes that non-spinning reserves are provided by resources flagged as off-line in the resource plan, while the previous analysis assumed that non-spinning reserves can be provided by resources that are physically off-line but flagged as on-line in the resource plan.

of their available energy is to manage the ramp capability of their portfolio. There are several factors that could contribute to fluctuations in un-offered energy.

- First, the fraction of responsive reserves satisfied with demand response has increased from less than one-third of the requirement in 2003 to approximately one-half in 2005. Thus, less of the on-line capacity is needed to satisfy the responsive reserves requirement, which allows more provide balancing energy.

- Second, wind capacity has risen over the period which tends to increase the amount of on-line capacity that QSEs set aside for portfolio balancing.

- Third, a larger share of OOME instructions were made in the down direction in 2004 compared with either 2003 or 2005. QSEs that receive upward portfolio balancing instructions at the same time as downward OOME instructions may find it difficult to meet both at the same time. This could lead some QSEs to set aside on-line capacity for portfolio balancing purposes.

- Fourth, QSEs reduced their practice of treating off-line units as quick-start resources that are available in real time in 2005 due to a change in market rules.[34] ERCOT had frequency control problems that it attributed to schedule control error caused by off-line units that could not start quickly enough. PRR 588 was created to restrict this practice to units that are capable of starting and ramping to maximum output in the ten minute ramping period between intervals. Generators must now actually demonstrate this capability to ERCOT.

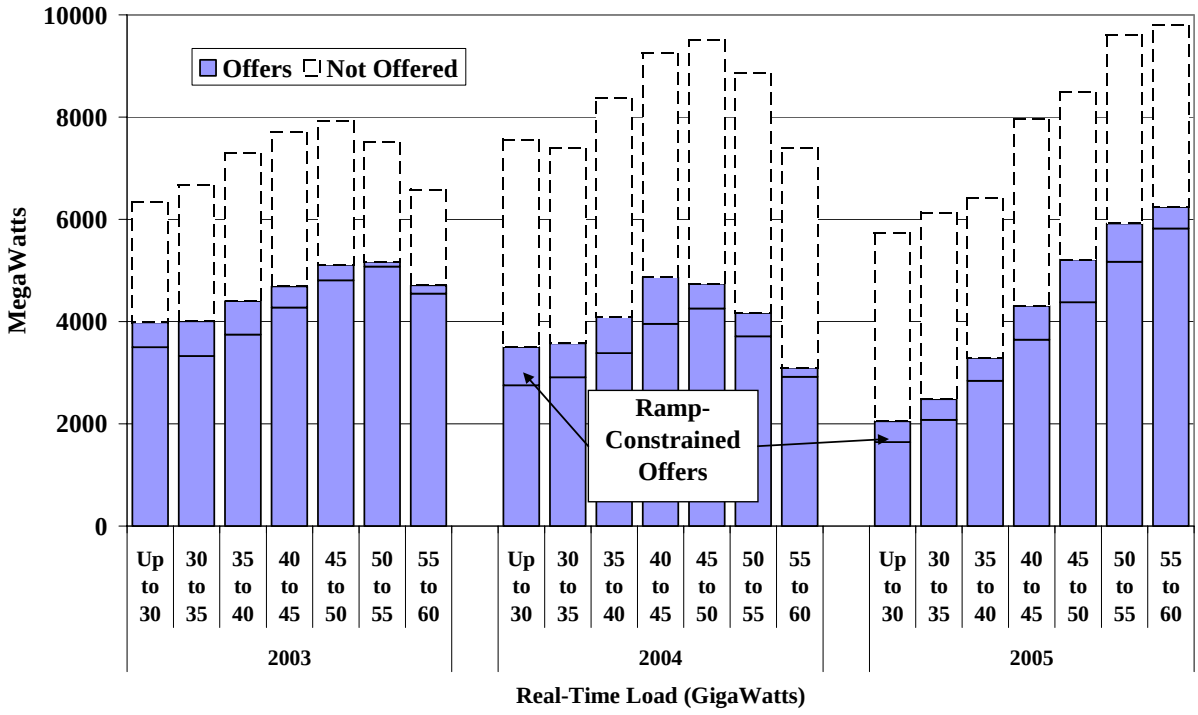
The sizable quantity of un-offered capacity is disturbing because it indicates that there is a substantial amount of energy that is not being efficiently utilized. It is of particular concern due to the more frequent price spikes that began to occur toward the end of 2004 and continued throughout 2005. It is also relatively common for the balancing energy market to experience shortages when on-line resources have the capability to increase their output. In 2005, the balancing energy market was short of up-balancing supply in 391 intervals, or more than once per day on average. When the balancing energy shortage is large enough to be a concern for reliability, the ERCOT operators give fleet OOME instructions to suppliers that did not offer all of their available capacity to the balancing market. This practice highlights the fact that significant amounts of capacity are not being utilized efficiently. As the surplus capacity in

[34] QSEs submit resource plans indicating which resources they plan to have on-line to satisfy their energy and ancillary service schedules plus any balancing energy deployments. The balancing market model is limited to giving deployment instructions that do not exceed the capacity of the QSE's resources that are flagged as on-line in the resource plan. QSEs are allowed to submit resource plans flagging as on-line resources that are actually off-line but capable of starting quickly.

ERCOT dissipates and it becomes more important to fully utilize the existing generating resources, these patterns will have much more serious implications.

Un-offered energy can also raise competitive concerns since a dominant supplier could possibly attempt to exercise market power by strategically withholding this energy from the balancing energy market. To investigate whether this has occurred, Figure 42 shows the same data as the previous figure, but arranged by load level for daily peak hours in each year.[35]

Figure 42: Balancing Energy Offers compared to Available Capacity
Daily Peak Load Hours – 2003 to 2005



The figure indicates that in 2003, the average amount of capacity available to the balancing market increased gradually up to 50 GW of load and then declined at higher levels. The pattern was similar in 2004, although the available capacity declined more substantially above 50 GW. In 2005, the available capacity continued to increase even at the highest load levels. Much of the rise in available capacity was due to the greater degree of under-scheduling in the summer of 2005 relative to the previous summers. In 2003 and in 2005, the fraction of available capacity

[35] More precisely, available capacity and balancing-up offers were ascertained for the peak hour of each day from 2003 to 2005. This data was then separated by load level and the average capacity and average balancing-up offers were calculated.

offered to the balancing market grew as load increased. In 2004, the fraction offered was more consistent across load levels.

The pattern of un-offered capacity shown in Figure 42 does not generally raise significant competitive concerns. If the capacity were being strategically withheld from the market, we would expect it to occur under market conditions most susceptible to the exercise of market power. Thus, we would expect more un-offered capacity under higher load conditions. However, the figure shows that portions of the available capacity that are un-offered are relatively consistently under all load conditions.

Aside from strategic withholding, there are several possible explanations for the large quantity of un-offered on-line and quick start capacity. First, the issues related to ramp rates discussed in the prior subsection can affect the offer levels. Currently, QSEs are able to submit one up-balancing ramp rate for its portfolio, although the previous subsection highlighted the fact that ramp capability tends to decrease as more of the offer is deployed. Thus, many QSEs may feel compelled to not offer slow ramping portions near the high sustainable limits of their resources. Moreover, 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 that would compelling it to dispatch its high-cost resources at a loss.

Second, it is very difficult to offer gas turbines in the balancing energy market effectively. The available capability in Figure 41 includes off-line gas turbines indicated as quick-start by the QSE in the resource plan. Although this practice became less common in 2005 due to the change in market rules discussed above, a substantial portion of the non-offered capacity is likely from these resources. The owners of gas turbines face significant challenges due to timing and minimum run-time considerations. An off-line gas turbine can start-up in response to a price spike in the balancing market, although there is no assurance that prices in subsequent intervals will support the continued operation of the gas turbine. This can cause a QSE with a gas turbine that is satisfying its portfolio instruction 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 rational that some suppliers do 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).

Lastly, the duct firing ranges of combined cycle units and steam turbines can also be difficult to offer in to the balancing market for several reasons. A supplier may incur "start-up costs" associated with operating in the duct firing range. Typically, generators have slower ramp rates in their duct firing ranges and may incur losses if a brief price spike is followed by relatively low prices. Many generators cannot operate in the duct firing range and provide regulation simultaneously. Thus, suppliers with duct firing capacity face many problems similar to those faced by the operators of off-line gas turbines.

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 43 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 three owners of generating capacity in ERCOT, whereas all other QSEs are included in the "small" suppliers category.

Figure 43: Balancing Energy Offers versus Available Capacity in 2005
Large and Small Suppliers -- Daily Peak Load Hours

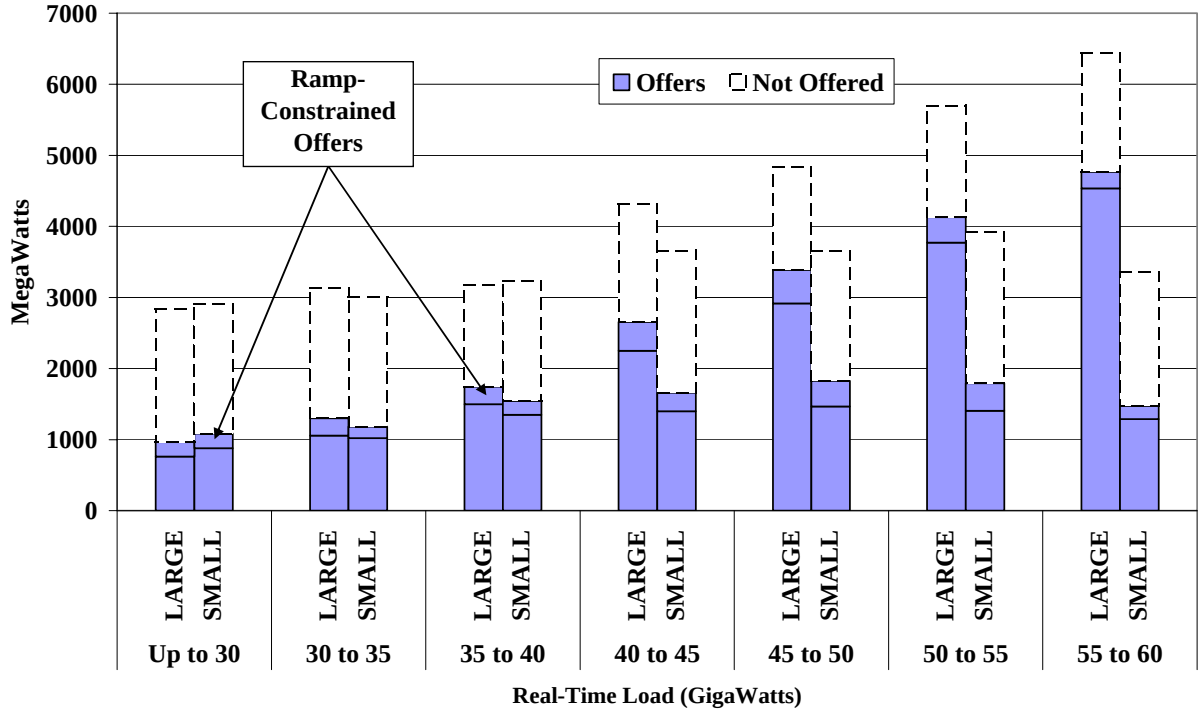
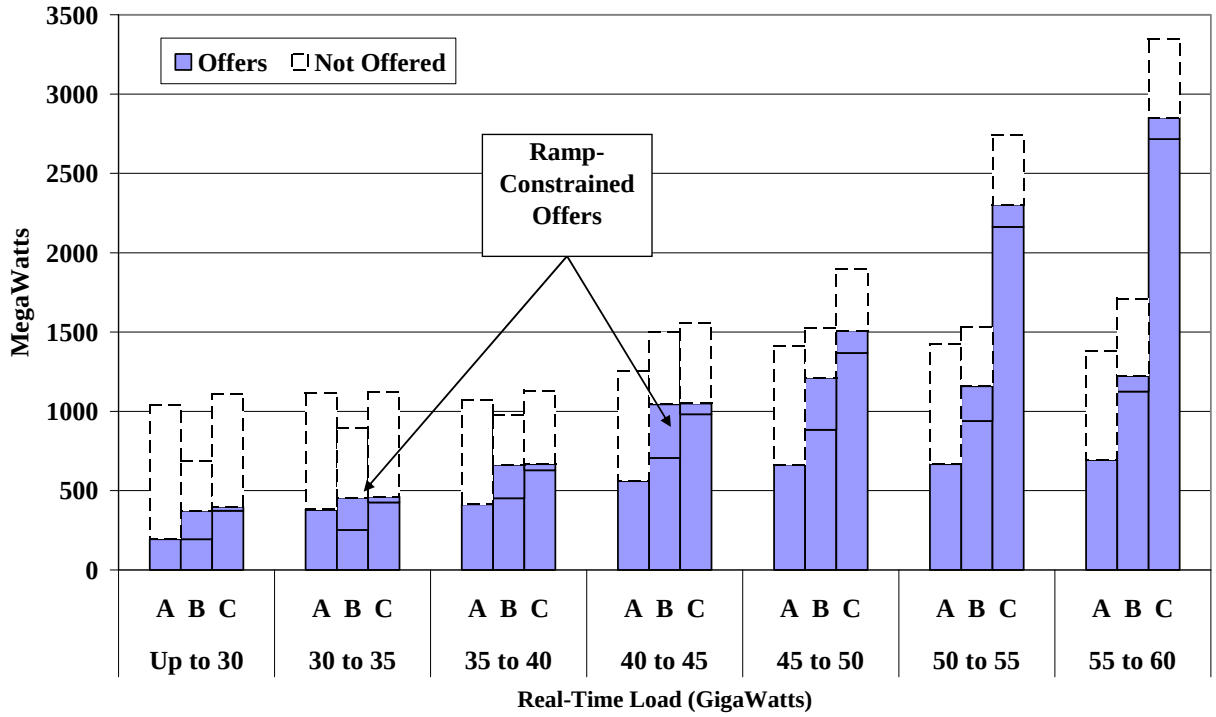


Figure 43 shows that the amount of available capacity in the portfolios of small suppliers (measured by the sum of the stacked bars) tends to be consistent across actual load levels. The portion of this available capacity that is offered in the balancing market is also relatively consistent across load levels. Conversely, the amounts of available capacity and offered capacity in the portfolios of large suppliers increase with the overall load level. At the lowest load levels, large participants offered an average of 1000 MW or 35 percent of their available capacity. At the highest load levels, large participants offered nearly 75 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.

Large participants are more likely than small participants to have market power and generally have the greatest incentive to withhold during periods of high demand. Furthermore, Figure 44 indicates that large participants hold a disproportionately large share of the available capacity during high demand periods, thereby increasing the ability to affect outcomes in the balancing energy market. However, we observe from the figure above that large participants offer their generation at a higher rate than small suppliers during the highest demand periods. Since small

suppliers generally do not have incentives to withhold, it is possible that the market design problems associated with ramping issues and portfolio bidding discussed earlier in this section affect small suppliers more than large suppliers. A more detailed analysis is needed before drawing conclusions about withholding behavior by either large or small suppliers. While investigating the specific conduct of large or smaller suppliers is beyond the scope of this report, the following figure summarizes the offer patterns of each of the largest three QSEs, individually. These are labeled as "A", "B", and "C" in Figure 44.

Figure 44: Balancing Energy Offers versus Available Capacity in 2005
Three Largest Suppliers -- Daily Peak Load Hours



Of the three suppliers, Company C generally had the largest amount of available energy, with approximately 1 GW at low to moderate load levels, rising above 3 GW at high load levels. Company A and Company B exhibited increasing but more consistent levels of available energy as was load rising. Company C and Company B both generally offered 70 to 85 percent of their available energy to the balancing market under high load conditions, although Company B's offers were more restricted by the portfolio ramp rate. As a result of ramp constraints, 30 to 50 percent of Company B's offered energy was not immediately dispatchable, however, the ramp constrained offers were relatively small at the highest load levels. Company A offered a smaller

share of its available energy than Company B and Company C, ranging from 20 percent at the lowest load levels to 50 percent under high load conditions.

If one of the three largest suppliers was withholding by not offering to the balancing market, we would expect them to offer less under the highest load conditions. Figure 44 does not indicate this pattern for any of the large suppliers. However, there are large amounts of un-offered capacity under all conditions by all suppliers that would tend to mask any withholding activity. Thus, without doing a more detailed analysis, it is impossible to draw definitive conclusions regarding whether any QSEs have strategically withheld capacity in some case by not offering it in the balancing energy market. However, the overall pattern is consistent with large suppliers offering more as load increases.

E. Resource Plan Changes

QSEs must have sufficient generation on-line to support their energy schedules and offers, and they are required to inform ERCOT about which resources they plan to use to satisfy their obligations. They do this by submitting resource plans at various points in the day-ahead and the operating day. While QSEs are expected to make their best effort to accurately forecast how they will operate their units, the resource plans are not financially binding and can be changed until shortly before real-time.[36] Resource plans are used by ERCOT in some of its reliability assessments before real-time and to make additional commitments to maintain reliability. Therefore, it is important for ERCOT to have accurate information in the resource plans that QSEs submit in order to avoid taking unnecessary and sometimes costly actions to maintain reliability.

It is important for QSEs to have the flexibility to incorporate new information prior to real time, such as demand forecast changes, generation and transmission outages, and other factors that suggest more or less resources will be needed in real-time. These factors can lead QSEs to significantly revise their resource plans after the day ahead. Under the current ERCOT market, however, there are other reasons why a participant may consistently provide unreliable information in its day-ahead resource plan, then revise the resource plan prior to real time when

[36] While resource plans are not financially binding, the real-time planned generation is used in the OOME payment formulas to determine the amount of megawatts deployed by the OOME instruction.

the balancing energy market is run. While it is possible for participants to submit unreliable information as part of a gaming strategy, they might also unintentionally submit unreliable information.

This section of the report analyzes the changes in the resource plans between the day ahead and real time and differences between the real-time resource plan and actual operation. Specifically, we evaluate units that are frequently committed out-of-merit or frequently dispatched out-of-merit. Such units receive additional payments from ERCOT and we investigate whether market participants may engage in strategies to increase the probability of receiving these payments.

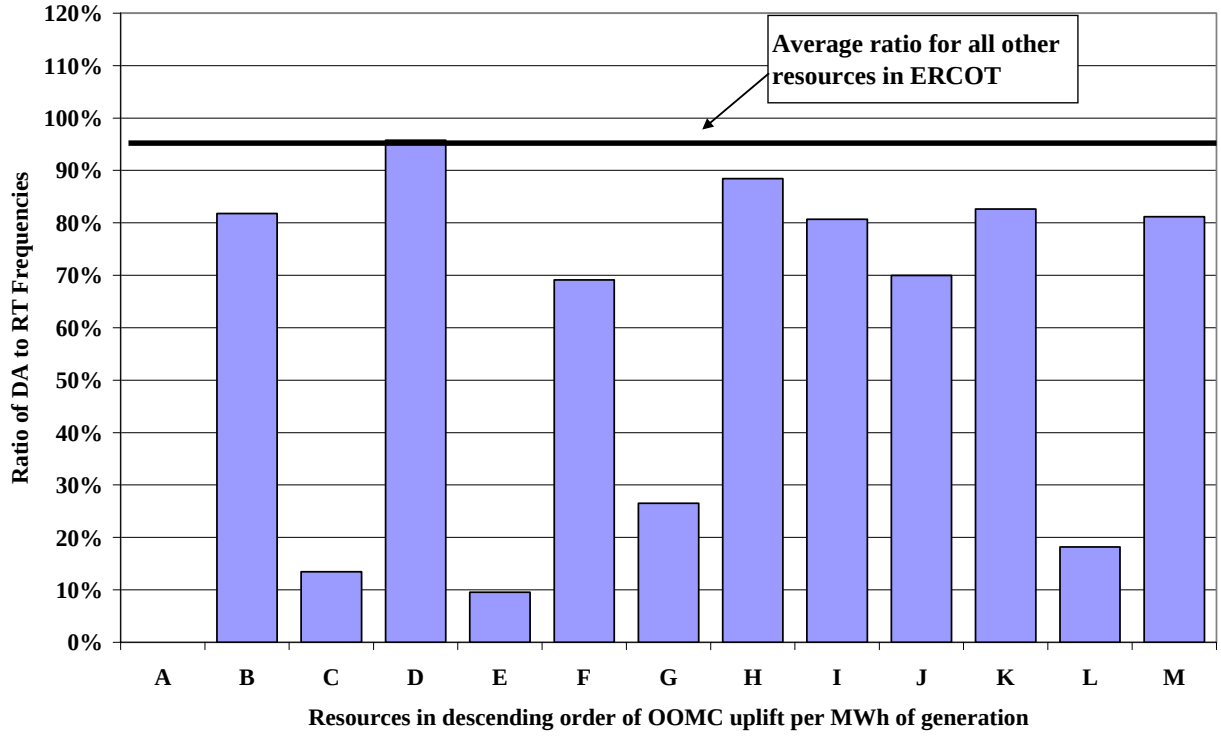
We first analyze the behavior of suppliers that are the primary recipients of payments by ERCOT for out-of-merit capacity. OOMC occurs when ERCOT instructs a unit that is not committed in the QSE's day-ahead resource plan to start in order to ensure sufficient capacity in real time to meet forecasted load and manage transmission constraints. When suppliers receive OOMC instructions, they receive payments from ERCOT that are designed to cover an estimate of the cost of starting the unit plus the cost of running at the minimum level. However, the OOMC unit retains any profits from sales above the minimum level into the balancing market. Thus, for units with significant commitment costs that are frequently committed out of merit, a supplier has the financial incentive to show the unit as uncommitted in the day-ahead resource plan to compel ERCOT to commit the unit. This supplier can subsequently commit the unit before real time if it is not OOMCed.

Because of the incentives presented by the OOMC process, we would expect suppliers that anticipate having units committed out-of-merit and that would benefit from the resulting payments to avoid showing the units as committed until after the out-of-merit commitments are announced. We examined the patterns of commitment for units that receive substantial OOMC payments. Figure 45 shows the ratio of day-ahead resource plan commitments to actual real-time commitments during 2005 for the 13 resources receiving the largest OOMC payments per MWh of production.[37] This should identify the resources with the largest incentives to engage in this strategy. Hours when the resources are under OOMC or OOME instructions are not

[37] 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 as well as resources that operated in-merit for fewer than 10 hours.

included in order to assess systematic changes made voluntarily by market participants. The units are shown in decreasing order of payments received from ERCOT on a per MWh basis—from $53 per MWh of generation across all hours (not just OOMC hours) for the units on the far left to $16 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 45: Ratio of Day-Ahead to Real-Time Resource Plan Commitments*
Frequent OOMC Resources – 2005



* Excluding hours when resources were under OOMC instructions or dispatched out-of-merit.

All of the resources shown in Figure 45 have ratios of less than 100 percent, ranging from 0 percent to 96 percent. In contrast, the average ratio for all other units is 95 percent, reflecting a much higher consistency between the day-ahead and real-time resource plans for the market as a whole. The results shown in this figure are consistent with the concern that some QSEs generally wait until after the OOMC process to commit units that are necessary for local reliability. This is consistent with our findings from 2003 and 2004.[38]

38 See 2003 SOM Report and 2004 SOM Report

For the resources shown in Figure 45, uplift payments for OOMC commitments are substantial enough to provide significant incentives to behave in ways that maximize the likelihood of receiving them. Figure 45 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 approach to address capacity insufficiency in the Protocols has several deleterious effects on the market. First, ERCOT 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. However, the QSE generally has the opportunity to modify its other commitments after it receives the OOMC instruction and often does so. Third, this conduct tends to make unreliable the information that ERCOT depends on to manage reliability. Ultimately, this can cause ERCOT to take a variety of costly actions, including making out-of-merit commitments that should not be necessary. These problems stem from the de-centralized process for unit commitment under the current market design, and underscore the reliability and efficiency benefits of the centralized commitment process that will be implemented with the nodal market re-design.

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.

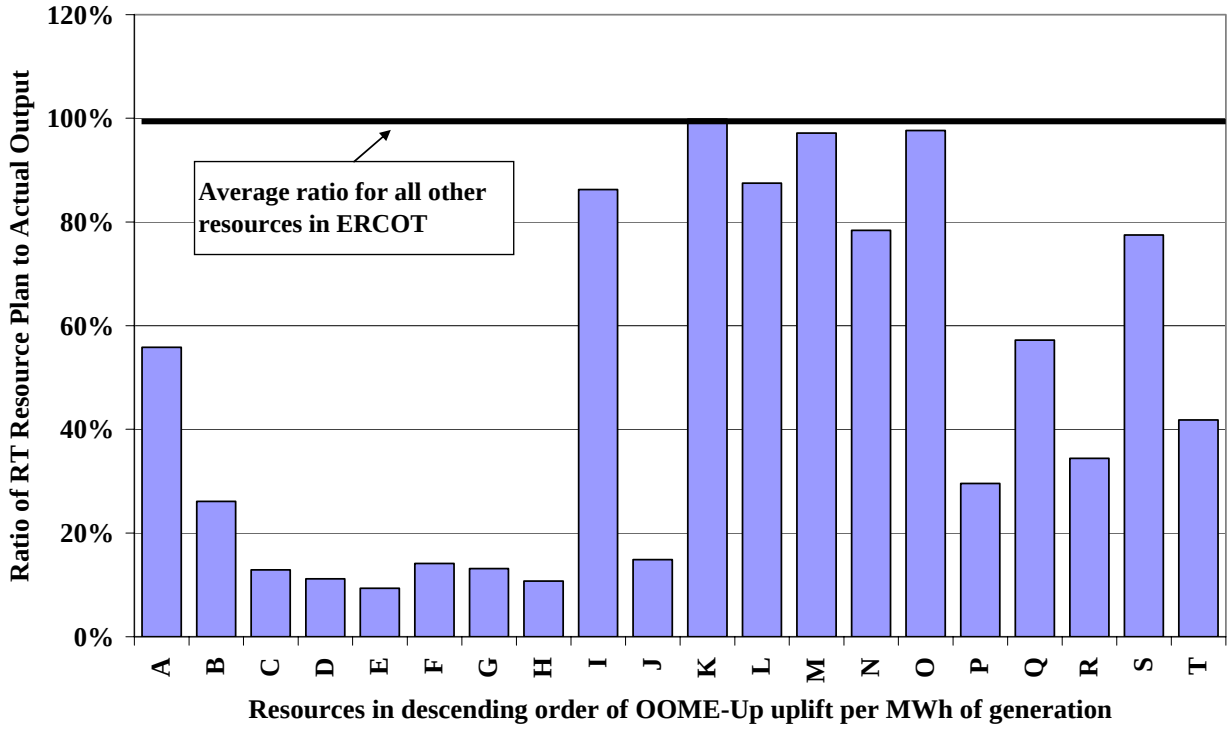
If a supplier is able to predict which of its units may be dispatched out-of-merit, it may under-schedule those units and overschedule other units in its portfolio.[39] Although this resource plan output may not be efficient, it can be effective at compelling an OOME instruction and the

[39] "Scheduling" in this context refers to the unit-specific planned generation in the QSEs' resource plans.

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 46 shows the ratio of real-time resource plan scheduled output to actual generation for the 20 units that received the highest average payments for OOME up per MWh of generation across all hours of 2005.[40]

Figure 46: Ratio of Real-Time Planned Generation to Actual Generation* Frequent OOME-Up Resources – 2005



* Excluding hours when resources were under OOMC instructions or dispatched out-of-merit.

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. The 20 resources

40 To focus on the most significant units, the analysis excludes resources where total uplift was less than $2 per kW-year of capacity or the average was less than $2 per MWh of generation as well as resources that operated in-merit for fewer than 10 hours.

shown in Figure 46 are presented in decreasing order of average payments, from $20 per MWh of generation across all hours for the unit on the far left to $2 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 20 resources shown in Figure 46, 11 have ratios of less than 50 percent while the rest have ratios between 50 and 100 percent. The other units in ERCOT had a weighted average ratio of 98 percent during the period, reflecting 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.

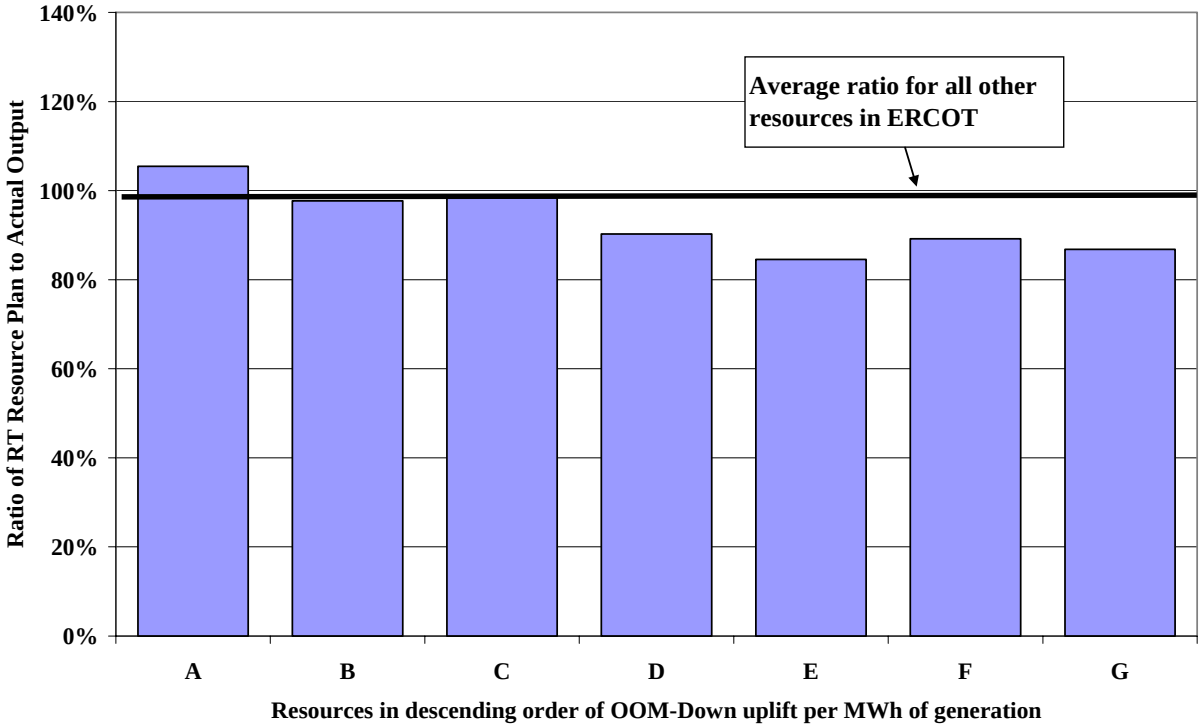
We next evaluate the incentives associated with providing OOME down. The incentives associated with rules for OOME down payments are the reverse of the incentives for OOME up payments. Since ERCOT pays units to reduce output from the real-time resource plan output levels, a supplier able to foresee the 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 47 shows the ratio of real-time resource plan output to actual generation for nine select resources that earned the highest average payments for providing OOME down (on per MWh basis) in 2005.[41] The figure shows the seven units that received the highest OOME down payments for their total production. The seven resources are shown in decreasing order of the average OOME down payments received per MWh of output, ranging from $2.01 per MWh on the far left to $1.01 per MWh on the far right. For comparison purposes, the figure also shows

[41] 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.

the generation-weighted average ratio of real-time resource plan output to actual generation for all other units.

Figure 47: Ratio of Real-Time Planned Generation to Actual Generation* Frequent OOME-Down Resources – 2005



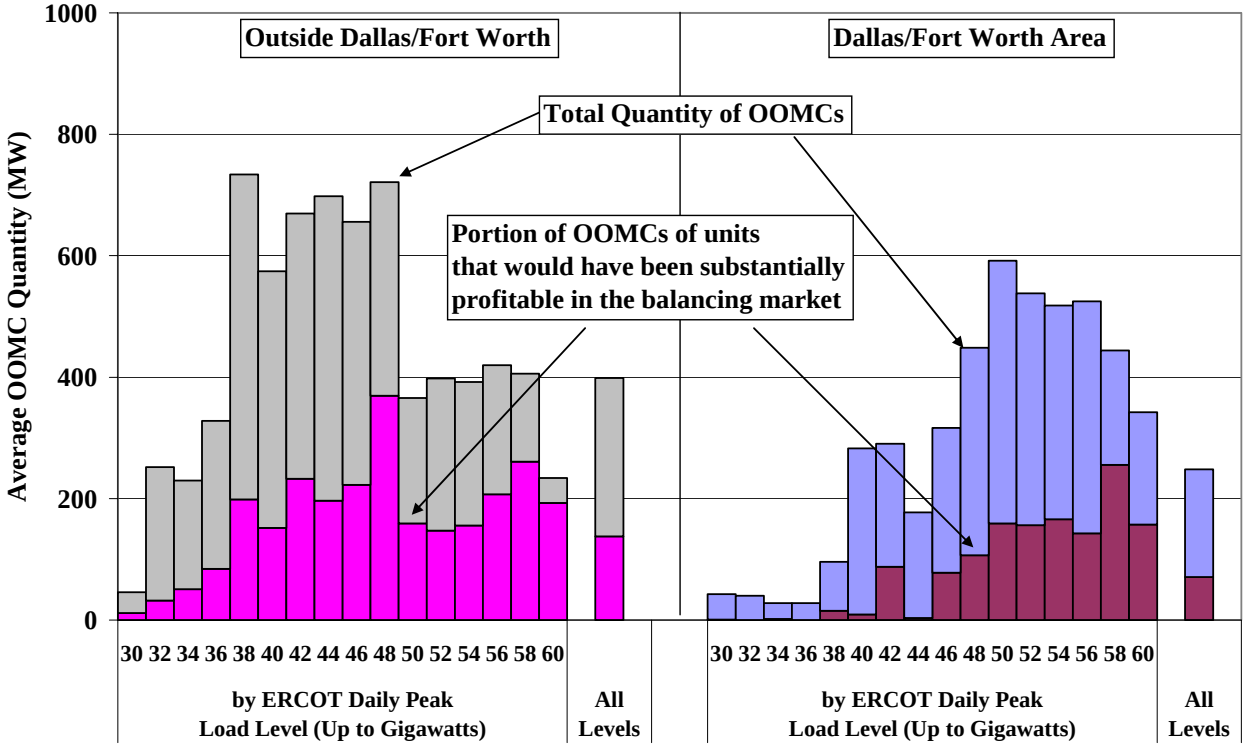
* Excluding hours when resources were under OOMC instructions or dispatched out-of-merit.

Only one of the seven of the resources shown in Figure 47 has a ratio above 100 percent, while the remaining six have ratios that range from 85 percent to 98 percent. This is in comparison with the average ratio exhibited by other units in ERCOT of 97 percent during the same period. The figure above reflects good consistency between the planned output level and actual generation for OOME down units. Thus, there is no indication that frequent OOME down units have systematically over-scheduled their resources to earn more OOME uplift.

Finally, we conducted a further analysis of the local congestion and out-of-merit commitment patterns. Historically, the majority of the OOMCs have been of capacity in the Dallas/Ft. Worth area, although it made up a smaller share in 2005. Figure 48 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.[42]

Figure 48: OOMC Supplied vs. ERCOT Load Level
Dallas/Fort Worth and Other Areas, 2005



This figure shows that 38 percent of ERCOT's out-of-merit capacity was located in Dallas/Ft. Worth in 2005, while the remainder was spread across the South, Houston, and West zones. The figure shows that as the demand in Dallas/Ft. Worth rises, operators must take more out-of-merit actions to maintain reliability. This differs from the pattern outside Dallas/Ft. Worth where a larger share of OOMC commitment was made when load was below 50 GW. This was primarily the result of brief spikes in out-of-merit commitments in Houston during May and early June, and in the South zone from October through December.

42 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.

Our previous analysis of resource plan changes between the day-ahead and real-time shown in Figure 48 indicates that units frequently committed out of merit are often voluntarily committed when ERCOT does not provide an OOMC instruction. This raises concerns about QSEs having the incentive to delay commitment decisions in order to garner OOMC payments. The results shown in Figure 48 indicate that both inside and outside Dallas/Ft. Worth, about one-third of resources receiving OOMC instructions would clearly have been economic for the QSEs to self-commit by a significant margin. The more profitable a unit was based on the balancing energy price, the easier it should have been for the QSE to predict this. Although, the results generally reinforce the concern that QSEs have delayed their commitment decisions in order to increase their uplift payments, there has been a reduction in OOMC instructions since 2003 which has likely diminished the incentives for QSEs to delay commitment decisions.

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. These concerns should be substantially addressed by the nodal market design, currently under development. A properly structured nodal electricity market should greatly reduce the need to commit and dispatch resources out of merit by incorporating the value of these resources in locational price signals. Such markets would substantially improve the efficiency of the management of local congestion.

III. DEMAND AND RESOURCE ADEQUACY

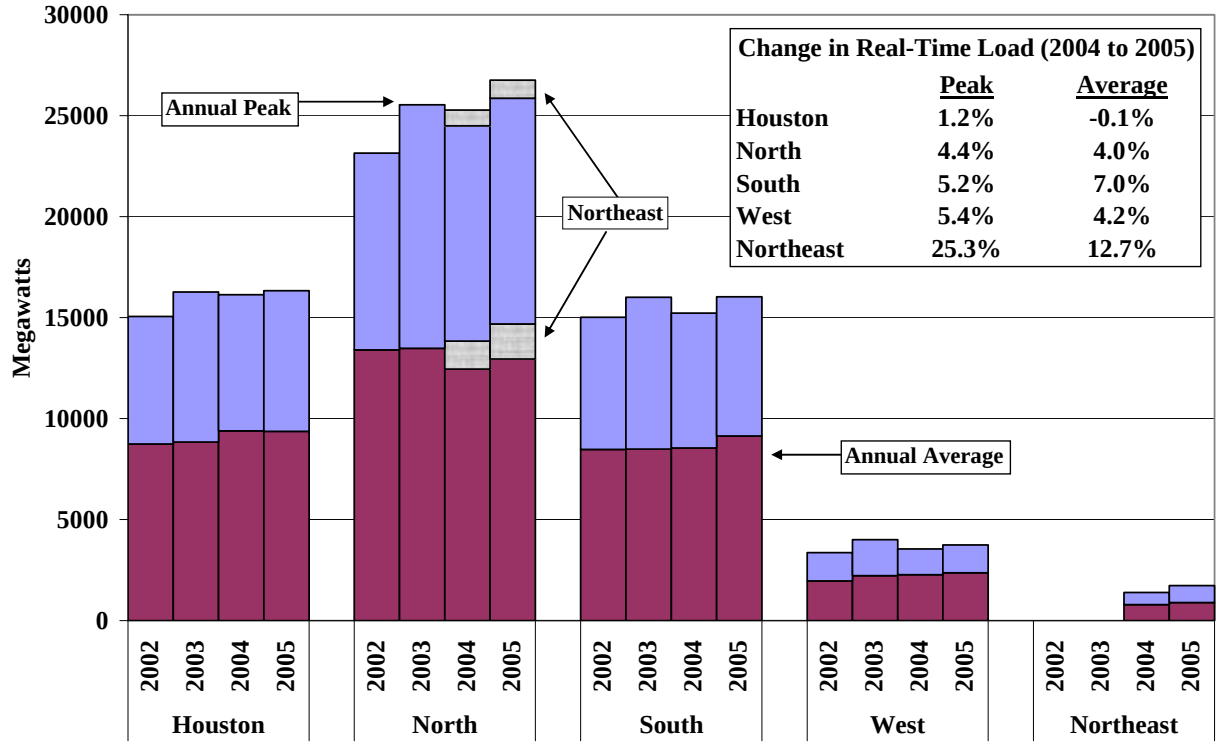
The prior sections of this report reviewed the market outcomes and provided analyses of a variety of factors that have influenced the market outcomes. This section reviews and analyzes the load patterns during 2005 and the existing generating capacity available to satisfy the load and operating reserve requirements.

A. ERCOT Loads in 2005

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 2005 are examined in this subsection and summarized in Figure 49.

This figure shows peak load and average load in each of the ERCOT zones from 2002 to 2005. It indicates that in each zone, as in most electrical systems, peak demand significantly exceeds average demand. The North Zone is the largest zone (about 37 percent of the total ERCOT load); the South and Houston Zones are comparable (with about 26 percent and 27 percent, respectively) while the West Zone and Northeast Zone are the smallest (with about 7 percent and 3 percent of the total ERCOT load). Figure 49 shows the annual non-coincident peak load for each zone. This is the highest load that occurred in a particular zone for one hour during the year, however, the peak can occur in different hours for different zones. As a result, the sum of the non-coincident peaks for the five zones was greater than the annual ERCOT peak load of 60,302 MW by approximately 1,350 MW in 2005.

**Figure 49: Annual Load Statistics by Zone*
2002 to 2005**

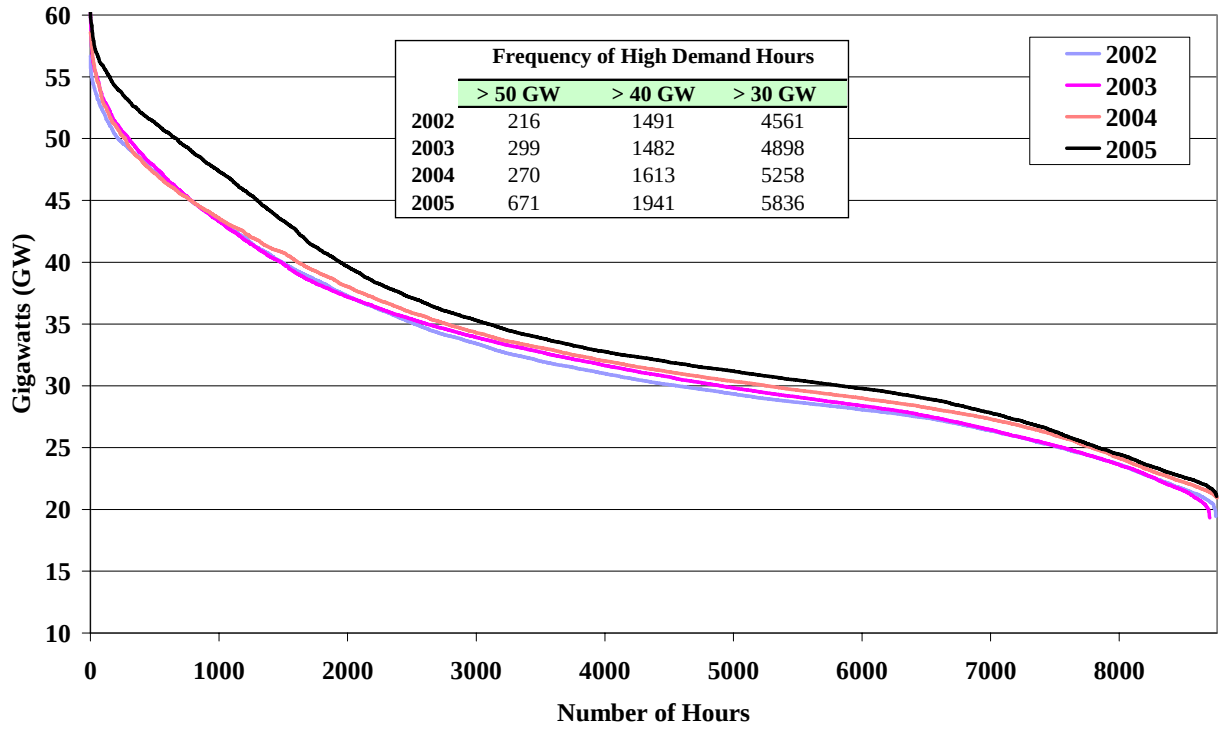


* The figure above is based on the load that SPD uses to schedule supply in the balancing energy market. This can differ from actual load in individual intervals.

No load statistics are shown for the Northeast Zone before 2004 because it was separated from the North Zone at the beginning of 2004. For comparison purposes, the Northeast Zone is also shown stacked with the North Zone in 2004 and 2005. The Northeast zone showed the most significant load growth, partly because several load busses were shifted from the North zone to the Northeast zone between 2004 and 2005.

To provide a more detailed analysis of load at the hourly level, Figure 50 compares load duration curves for each year from 2002 to 2005. 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 2005, the highest load hours occurred in the summer months, particularly in August.

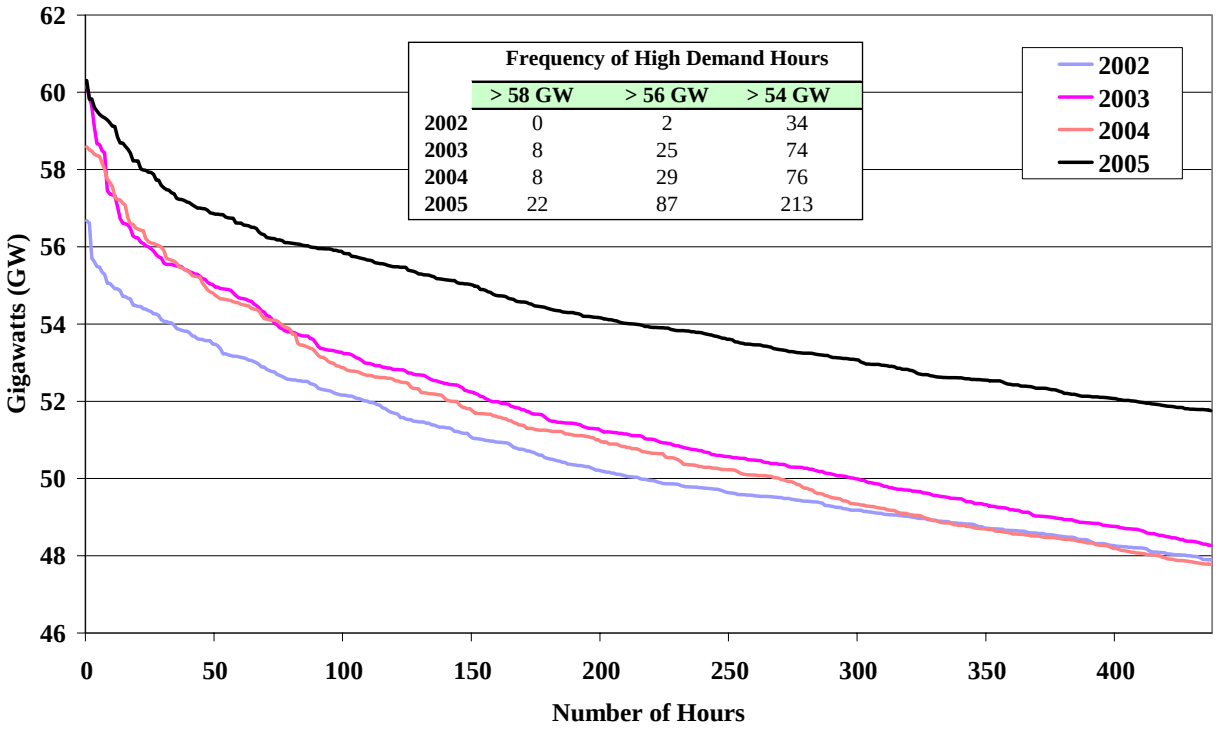
Figure 50: ERCOT Load Duration Curve
All Hours – 2002 to 2005



As Figure 50 shows, the load duration curve for 2005 lies above the ones for the previous three years. Load increased more from 2004 to 2005 than it did in the previous two years, particularly in the highest 2000 hours due to relatively high weather-dependent loads. In 2005, there were 20 percent more hours when load exceeded 40 GW and more than twice as many hours when load exceed 50 GW than in 2004.

To better show the differences in the highest-demand periods between years, Figure 51 shows the load duration curve for the top 5 percent of hours with the highest loads. It shows that while load increased modestly in each year from 2002 to 2004, the increase from 2004 to 2005 was much larger. Load exceeded 58 GW in 22 hours in 2005 and 8 hours in 2003 and 2004. In 2002, demand was not higher than 58 GW in any hour. The same pattern prevailed at lower load levels with 2005 demand being considerably higher than in previous years.

Figure 51: ERCOT Load Duration Curve
Top Five Percent of Hours – 2002 to 2005

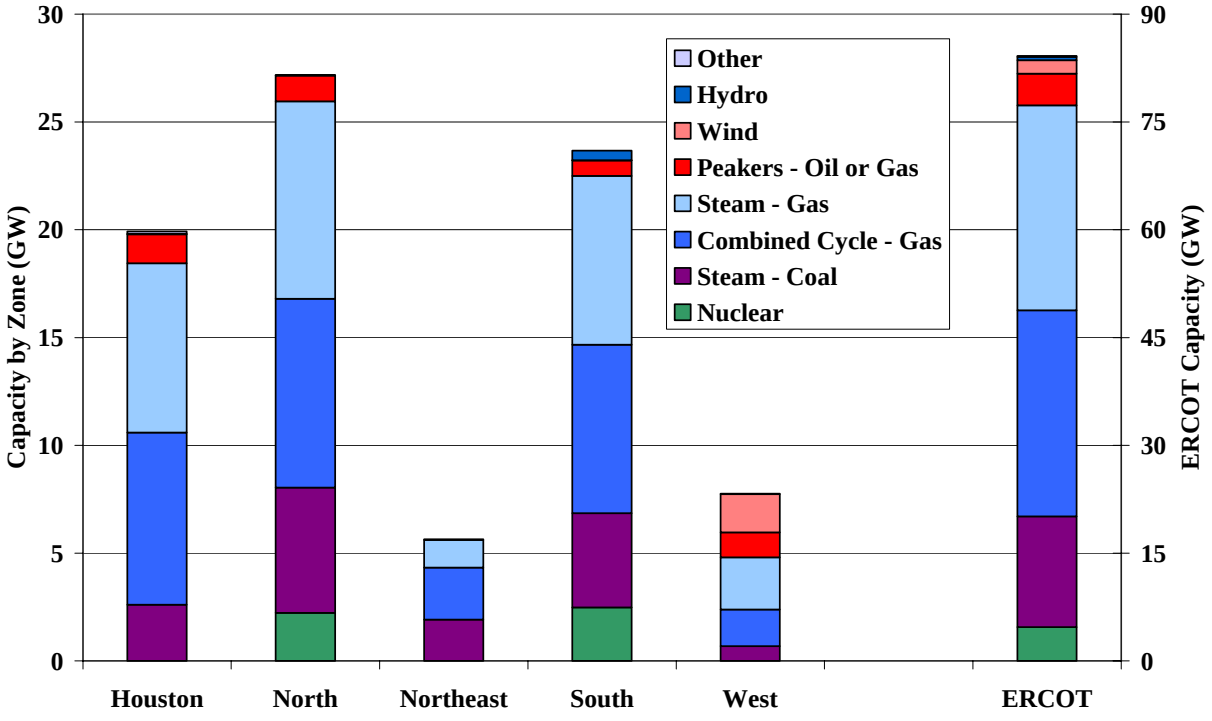


This figure also shows that the peak load in each year was roughly 15 to 25 percent greater than the load at the 95th percentile of hourly load. For instance, in 2005, the peak load value was over 60 GW while the 95th percentile was lower than 52 GW. This is typical of, and even somewhat flatter than, the load patterns in most electricity markets. This implies that a substantial amount of capacity, more than 8 GW, is needed to supply energy in less than 5 percent of the hours. 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 52 shows the installed generating capacity by type in each of the ERCOT zones, including switchable capacity located outside ERCOT.

Figure 52: Installed Capacity by Technology for each Zone 2005



This figure shows that there is some nuclear capacity in both the North and South Zones, while lignite coal is also a significant contributor in ERCOT. However, the primary fuel in all five zones is natural gas (or sometimes oil) -- accounting for 73 percent of generation capacity in ERCOT as a whole, and 86 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. These new installations of capacity have not changed the overall mix significantly since the generators that have gone out of service during this period were primarily gas-fired steam turbines. These new units have increased the gas-fired share of installed capacity gas even though some older gas-fired units have been retired.

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 are very few hours when ERCOT load drops as low as 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

approximately 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 within separate control areas. The North Zone accounts for 32 percent of capacity, the South Zone 28 percent, the Houston Zone 24 percent, the West Zone 9 percent, and the Northeast Zone 7 percent. The North Zone and Houston are importers of power, while the Northeast Zone exports significant quantities because it has over two times more generation than its peak zonal load. Because large amounts of power flow out of the South Zone into the North Zone and Houston, the South-to-North CSC and the South-to-Houston CSC experienced the greatest amounts of congestion during 2005.

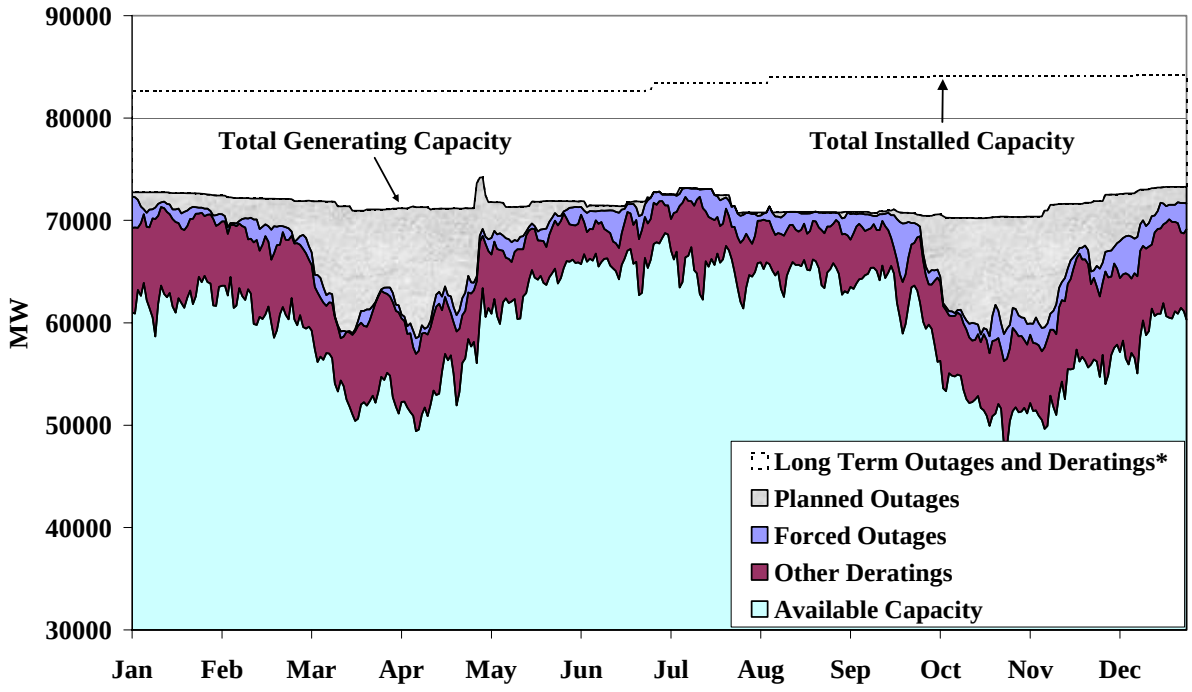
1. Generation Outages and Deratings

Figure 52 in the prior subsection shows that installed capacity far exceeds the annual peak load plus ancillary services requirements in ERCOT. This might suggest that the adequacy of resources is not a concern in ERCOT in the near-term, although 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 the installed capability of a generating resource 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 53 below shows a breakdown of total installed capability for ERCOT on a daily basis during 2005. 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, (c) short-term forced outages, (d) other short-term deratings, and (e) available and in-service capability.

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. Long-term deratings can occur for several reasons. First, some of this capacity may be out-of-service for extended periods due to maintenance requirements. Second, if their owners predict that wholesale market prices will not be sufficiently high to justify the periodic costs required to keep them available, some units may go out-of-service temporarily. Third, the owners of some cogeneration plants routinely use steam output to support their processes rather than generate electricity. 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 53: Short and Long-Term Deratings of Installed Capability**
2005



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

** Switchable capacity is included under installed capacity in this figure.

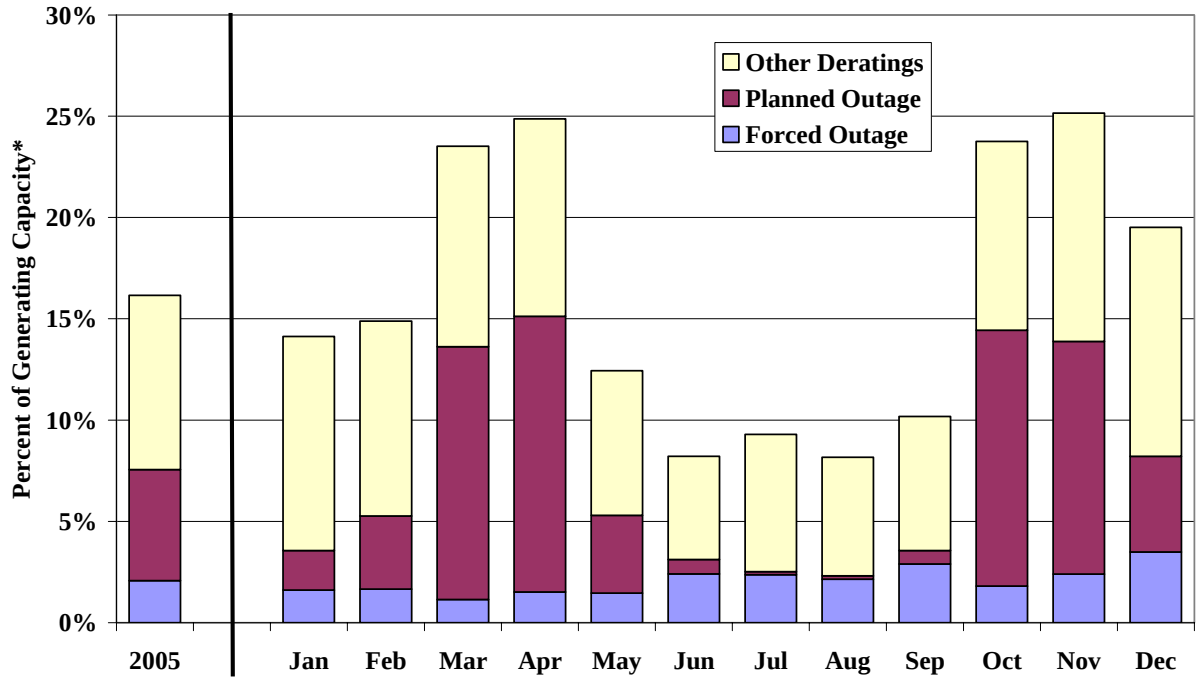
Figure 53 shows that installed capacity, including mothballed and switchable capacity, rose from 82 GW at the beginning of 2005 to 84 GW at the end of 2005. This increase is due to several new generators coming on-line although it was diminished by several retirements. The figure shows that the long-term outages and deratings fluctuated somewhat but generally grew from 10

GW at the beginning of 2005 to 11 GW at the end of the year. The long-term outages and deratings also include over 5 GW of mothballed capacity.[43] These classes of capacity can be made available if market conditions become tighter as load rises.

As expected, short-term planned outages are relatively large in the spring and fall, decreasing to close to zero during the summer. Available in-service capacity fluctuated between 48 GW in October and 68 GW in July. The peak hour for the year required 60.3 GW to satisfy ERCOT's energy requirements and an additional 3 GW for operating reserves and regulation-up requirements, resulting in surplus capacity of approximately 2 GW on that day.

The next analysis focuses specifically on the short-term forced outages and other short-term deratings. Figure 54 shows the average magnitude of the outages and deratings lasting less than 60 days for the year and for each month during 2005.

Figure 54: Short-Term Outages and Deratings*
2005



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

43 See "Report on the Capacity, Demand, and Reserves in the ERCOT Region," June 2005.

Figure 54 shows that total short-term deratings and outages were as large as 25 percent of installed capacity in the spring and fall, and dropped below 8 percent for much of the summer. Most of this fluctuation was due to anticipated planned outages, which ranged as high as 11 to 14 percent of installed capacity during March, April, October, and November. Short-term forced outages occurred more randomly, as would be expected, ranging between 1 percent and 4 percent of total capacity on a monthly average basis during 2005. 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 54 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.

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 deratings due to ambient temperature conditions, cogeneration uses, and other factors described above. Furthermore, suppliers may delay maintenance on components such as boiler tubes, resulting in reduced capability. Because these 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 2005 and as high as 11 percent in other months.

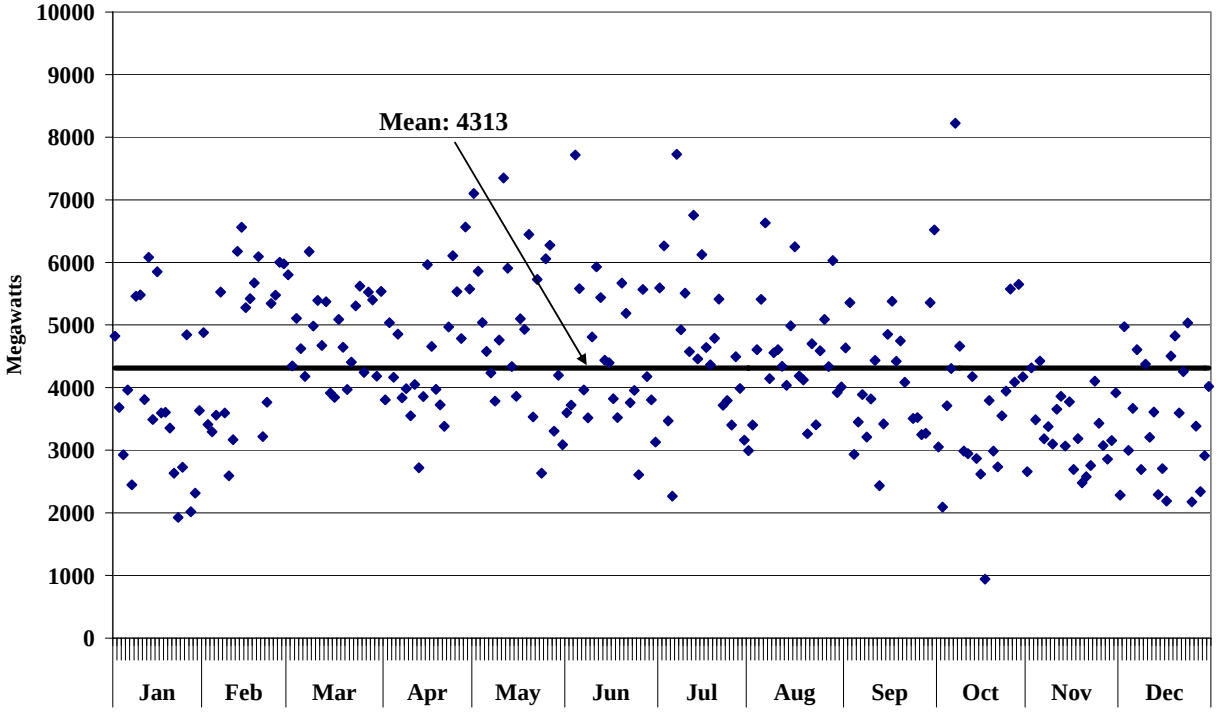
2. Daily Generator Commitments

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[44] 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. 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 55 plots the excess capacity in ERCOT during 2005. 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 55: Excess On-Line and Quick Start Capacity During Daily Peaks on Weekdays -- 2005



44 For the purposes of this analysis, "quick-start" includes (i) units that ERCOT has identified as capable of starting-up and reaching full output after receiving a deployment instruction from the balancing energy market, and (ii) off-line simple cycle gas turbines that are flagged as on-line in the resource plan with a planned generation level of 0 MW.

Figure 55 shows that the excess capacity in ERCOT was significant during 2005. The levels rarely fell below 2 GW on any day and sometimes exceeded 7 GW. During the peak load day in 2005 (on August 23), there were 3,407 MW of excess capacity available. The excess capacity averaged 4,313 MW, which is approximately 15 percent of the average load in ERCOT. As explained above, some of this excess capacity reflects the fact that it can be economic to commit steam units or combined cycle units to serve the peak load even when quick-start peaking resources are available. However, the off-line quick-start resources reflect a minority of the excess shown in the figure. 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 result in higher than necessary production costs.

In 2005, the average amount of excess committed capacity was approximately 35 percent less than in 2004. At least two factors contributed to this reduction. First, as mentioned in the previous section, the practice of QSEs treating off-line units as quick start became much less frequent in 2005 due to a change in market rules. PRR 588 was created to restrict the practice to units that could demonstrate to ERCOT the capability to start in the necessary timeframe. This reduced the amount of excess capacity from off-line resources that might have participated in the balancing market. Second, ERCOT committed fewer resources in 2005 via OOMC instructions which tend to result in over-commitment.

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 a day-ahead energy market with Security Constrained Unit Commitment ("SCUC"), under the nodal market design planned for implementation in 2009, 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").

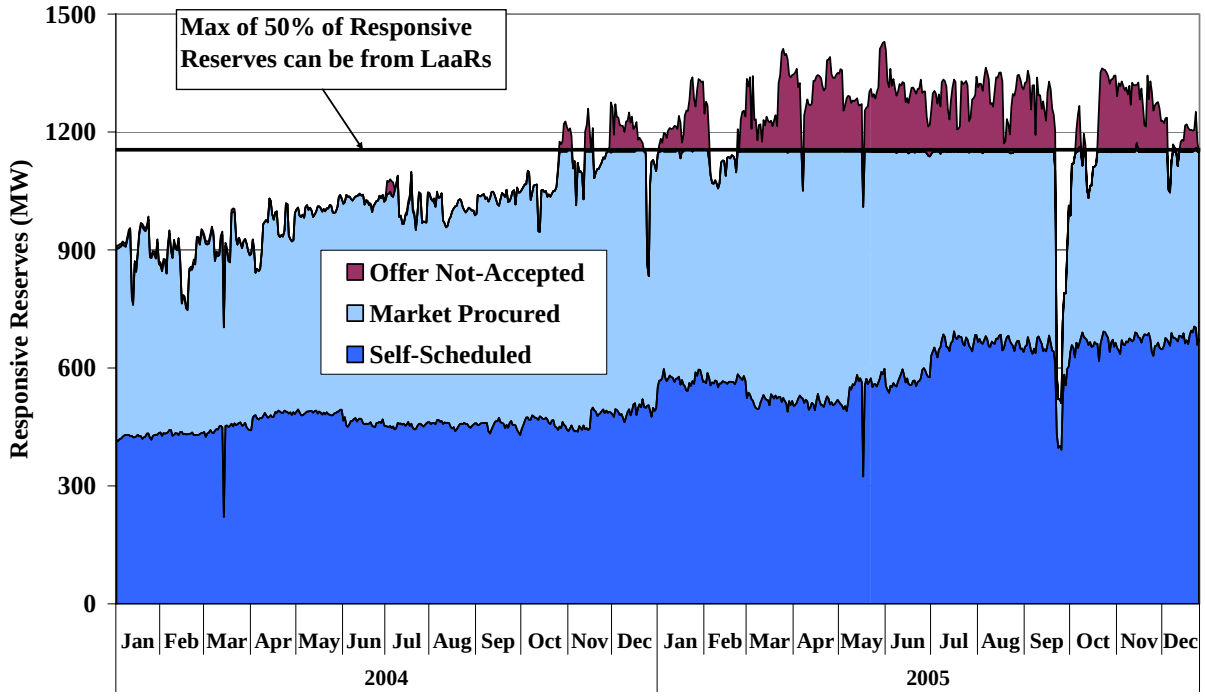
ERCOT allows LaaRs that are qualified to offer responsive reserves and non-spinning reserves into the day-ahead ancillary services markets. Those that are qualified can also offer blocks of energy in the balancing energy market. LaaRs providing up balancing energy must have telemetry and must be capable 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) are not currently able to provide regulation service.[45]

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 some LaaRs, however, they are not qualified to provide reserves or regulation service.

As of December 2005, 92 resources totaling 1,835 MW of capability were qualified as LaaRs. These resources regularly provided reserves in the responsive reserves market, but never participated in the balancing energy market and only a very small portion participated in the non-spinning reserves market. There were no BULs registered with ERCOT in 2005. Figure 56 shows the amount of responsive reserves provided from LaaRs on a daily basis in 2005.

[45] PRR 307 was adopted in 2002 to allow load resources to provide regulation if they have the capability.

Figure 56: Provision of Responsive Reserves by LaaRs
Daily Average – 2005



The high level of participation by demand response sets ERCOT apart from other operating electricity markets. Figure 56 shows that the amount of responsive reserves provided by LaaRs gradually increased from about 900 MW at the beginning of 2004 to an average of 1,133 MW in 2005. The majority of this increase was procured through self-provision and bilateral agreements rather than the ERCOT administered auction. Currently, LaaRs are permitted to supply up to 1,150 MW of the responsive reserves requirement. In 2005, it became commonplace for the 1,150 MW restriction to limit the set of demand resources that could provide responsive reserves. This has highlighted a flaw with the way that the ancillary services auction selects demand resources to provide responsive reserves.

The auction ranks responsive reserves providers according to their offer price from lowest to highest.[46] The auction goes up the offer stack until it reaches the 2,300 MW required quantity of

46 In October 2005, ERCOT began to use a simultaneous clearing model for regulation up, regulation down, responsive reserves, and non-spinning reserves. This selection mechanism is conceptually similar since resources are selected in merit order. However, a resource with a low-priced responsive reserves offer may be selected to provide another product, such as regulation up, if the reduced cost of the other product exceeds the added cost of not using the resource to provide responsive reserves. In this case, the clearing price for responsive reserves is the marginal cost to the system of meeting the reserves requirement. This is always equal to the marginal reserves provider's offer price plus the opportunity cost of not providing an

reserves. However, if the auction reaches the 1,150 MW limit before meeting the 2,300 MW requirement, the offers of any additional LaaRs cannot be used and are discarded. In such cases, the marginal generator resource sets the clearing price for responsive reserves at a level that exceeds the offer prices of some of the unaccepted offers from LaaRs.

This mechanism for selecting providers and determining clearing prices for responsive reserves is inefficient and leads to excessive reliability costs for consumers. Routinely, the quantity of LaaRs willing to supply responsive reserves at the clearing price exceeds the demand for this service (i.e. 1,150 MW). When supply exceeds demand for a product at the prevailing price, it should cause the price of the product to decrease until the market reaches a level where the supply equals demand. Under the current market design, there is no mechanism for this to happen since there is only one price for all responsive reserves. Since the Protocols limit the amount of responsive reserves that can be provided by LaaRs, the price of reserves provided by LaaRs should clear below the price of reserves provided by synchronized generators.

The design of this market encourages inefficient behavior by QSEs that want to sell responsive reserves from their demand resources. Under current market conditions, the clearing price for responsive reserves is usually set by a generator. In order to be selected, it is not sufficient for LaaRs to submit an offer price that is below the clearing price. The LaaR's offer must also be included among the lowest priced 1,150 MW of LaaRs. This gives QSEs an incentive to offer LaaRs at arbitrarily low (even negative) prices. Indeed, at the November 2005 meeting of the Wholesale Market Subcommittee, ERCOT reported that for the previous market day "the LaaR portion of the bid stack is ranging from $0 to -$17,500." Under these incentives, competition does not lead to having the most efficient resources provide responsive reserves. This also raises the concern that a negative LaaR offer could set the responsive reserves clearing price in the event that 1,150 MW of generators are bilaterally scheduled for reserves. In this unlikely event, LaaRs might receive large invoices to provide reserves, raising potential credit issues.

To improve the efficiency of responsive reserves pricing and incentives for suppliers, we recommend that ERCOT set separate prices for the two types of responsive reserves. The best way to accomplish this would be by having two responsive reserves constraints in the ancillary

alternate product in the auction.

services auction: (i) that the responsive reserves procurement (including bilateral schedules) be greater than or equal to 2,300 MW and (ii) that the responsive reserves procurement from LaaRs (including bilateral schedules) be less than or equal to 1,150 MW. The clearing price paid to generators would be equal to the shadow price of the first constraint only, while the clearing price paid to LaaRs would be equal to the shadow price of the first constraint minus the shadow price of the second constraint.

Under this proposal, whenever the 1,150 MW limit on LaaRs providing responsive reserves was binding, the clearing price for responsive reserves from LaaRs would be determined by the offer of the marginal LaaR. Whenever the 1,150 MW limit did not affect the selection of resources (i.e. the shadow price of the second constraint equals $0), the clearing prices would be identical for both types of responsive reserves providers. This recommendation would likely require changes to the ancillary services market clearing engine software.

Alternatively, this recommendation could be implemented in a much less costly manner that does not involve changes to the ancillary services auction software. This would involve a post-process that determines the clearing price for responsive reserves provided LaaRs. In hours when fewer than 1,150 MW of responsive reserves are scheduled to be provided by LaaRs, the clearing price would be the same for generators and LaaRs. However, in hours when 1,150 MW are to be provided by LaaRs, the clearing price would be set by the lesser of: (i) the clearing price for generators and (ii) the lowest-priced unaccepted LaaR offer (which would be higher than the highest-priced accepted LaaR offer). This simpler mechanism is likely to produce the same results as the more ideal solution with less expense. This is similar to a recommendation made by a Special Task Force of the Demand Side Working Group that was endorsed by the Wholesale Market Subcommittee.[47] The PRR has not yet been written for this recommendation.

Although LaaRs are active participants in the responsive reserves market, they did not offer into the balancing energy or regulation services markets and they averaged just 19 MW of non-spinning reserves sales. This is not surprising because the value of curtailed load tends to be very high, and providing responsive reserves offers substantial revenue with very little probability of being deployed. In contrast, providing non-spinning reserves introduces a much

[47] See minutes of the March 22, 2006 Wholesale Market SubCommittee meeting

higher probability of being curtailed. Participation in the regulation services market requires technical abilities that most 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, non-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 model 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, all other constraints not defined as zonal constraints (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.

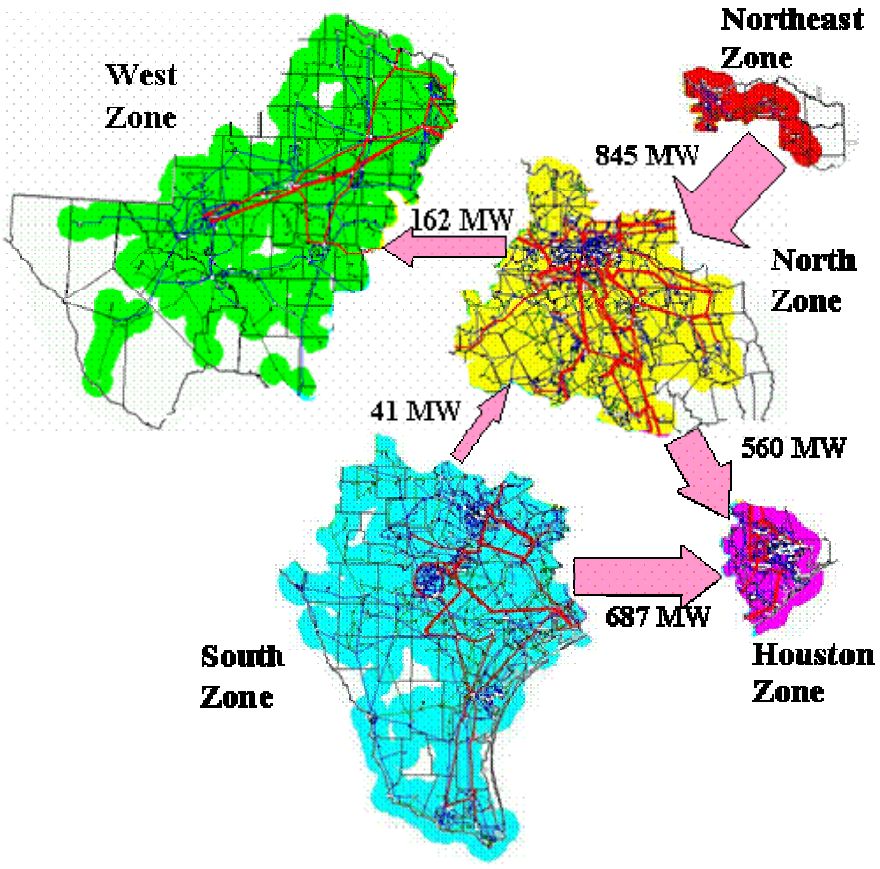
A. Electricity Flows between Zones

In 2005, there were five commercial pricing zones in ERCOT: (a) the North Zone, (b) the West Zone, (c) the South Zone, (d) the Houston Zone, and (e) the Northeast Zone, which was created in 2004 by dividing the North Zone. From year-to-year, slight adjustments are sometimes made to the boundaries of the commercial pricing zones, but the vast majority of customers remained in the same zone from 2004 to 2005. ERCOT operators use the SPD software to dispatch 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 five zone-based locations and six transmission interfaces. These six 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 utilize 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 2005.

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 are an indication of 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. In particular, it discusses the impact on congestion management of adding the new North-to-West CSC. Figure 57 shows the average SPD-modeled flows over CSCs between zones during 2005. A single arrow is shown for the modeled flows of both the North to West and West to North CSCs.

Figure 57: Average SPD-Modeled Flows on Commercially Significant Constraints During All Intervals in 2005



Note: In the figure above, CSC flows are averaged taking the direction into account. For instance, if one hour has a North to West flow of 100 MW, and a second hour has a West to North flow of 200 MW, the average hourly flow would be 50 MW from the West to North. This treats the North to West flows in the first hour as negative for averaging purposes.

Figure 57 shows the five ERCOT geographic zones as well as the six CSCs that interconnect the zones: (a) the West to North interface, (b) the South to North interface, (c) the South to Houston interface, (d) the Northeast to North interface, (e) the North to Houston interface, and (f) the newly created North to West interface. Since CSCs are defined in a single direction, the West to North interface, which existed prior to 2005, could not be used to manage the flows *into* the West zone, forcing ERCOT to do so using local congestion management procedures. Thus, the North to West interface was created to manage the increasingly frequent congestion into the West zone in a market-based way. Based on SPD modeled flows, Houston is a significant importer while the Northeast Zone and the South Zone export significant amounts of power.

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")[48] in relation to the CSC. In reality, the generators within each zone can have widely varying effects on the flows over a CSC. To illustrate this, we calculated flows that would occur over the CSC using actual generation and actual generation shift factors and compared this to flows calculated using actual generation and zonal average shift factors. Table 2 shows this analysis. The flows over the North to West CSC are not shown separately in the table below since they are equal and opposite the flows for the West to North CSC.

The first column in Table 2 shows the average flows over each CSC calculated by SPD. The second column shows the average flows over each CSC 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

[48] 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.

that the actual generation levels result in higher calculated flows on each CSC except the South to North, West to North, and South to Houston CSCs, where calculated flows are lower.

Table 2: Average Calculated Flows on Commercially Significant Constraints Zonal-Average vs. Unit-Specific GSFs – 2005

CSC 2005	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	-162	-176	*-14*	-195	*-20*
South-North	41	-14	*-55*	11	*26*
South-Houston	687	646	*-41*	884	*238*
North-Houston	560	580	*20*	492	*-89*
NorthEast-North	845	855	*10*	874	*19*

The fourth column in Table 2 reports the average flows over each CSC calculated using unit-specific 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 measure the inaccuracy caused by treating each unit within a particular zone as having identical impact 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 2 shows that the unit-specific GSFs increased the calculated flows on the South-Houston interface by 238 MW and reduced the calculated flows on the North to Houston CSC by 89 MW each. These differences are sizable and are generally 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. Using generation-weighted shift factors for load rather than load-weighted shift factors can cause significant differences between SPD flows and actual flows. For instance, SPD flows for the South to North interface will be approximately 188

MW higher than actual flows during summer conditions due to this simplification.[49] However, this impact of this assumption is diminished by the fact that loads are not used to manage transmission constraints in real-time. The use of simplified generation-weighted shift factors prevents the SPD model from efficiently assigning the costs of interzonal congestion. In the long run, the use of generation-weighted shift factors for loads systematically biases prices, so that buyers in some zones pay too much (e.g. the North Zone), and others pay too little (e.g. the South Zone).

In order to effectively manage interzonal congestion, it is important for SPD to accurately model the major constrained transmission interfaces between zones. In 2005, the six CSCs modeled by SPD did not include all significant interfaces between zones. Sizeable quantities of power were transported on transmission facilities not modeled by SPD. Table 3 summarizes the actual net imports into each zone compared to SPD modeled flows from 2003 to 2005.

Table 3: Actual Net Imports vs. SPD-Calculated Flows on CSCs 2003 to 2005

Year	Zone	Actual Net Imports	SPD Flows on CSCs
2003	**Houston**	1796	565
	North	-507	191
	South	-1213	-702
	West	-76	-54
2004	**Houston**	2479	1265
	North	867	264
	NorthEast	-2116	-858
	South	-1531	-800
	West	304	129
2005	**Houston**	2596	1247
	North	660	164
	NorthEast	-2138	-845
	South	-1501	-728
	West	386	162

[49] The annual planning study used by ERCOT to forecast transmission capability prior to the 2005 annual Transmission Congestion Rights auction calculates this effect to be 188 MW during summer peak conditions.

Table 3 summarizes the differences between average SPD-calculated flows and average actual flows into each zone. These differences can be attributed to three factors. First, the use of zonal average GSFs, rather than resource-specific GSFs, by SPD to model generators can cause the SPD-calculated flows on a particular CSC to be substantially different from the actual flows. Second, the use of generation-weighted shift factors to model load causes systematic differences between SPD flows and actual flows. For instance, SPD generally underestimated flows on the Northeast-North CSC because of the difference between load-weighted and generation-weighted shift factors, accounting for a significant chunk of the difference between SPD flows and net exports from the Northeast Zone. However, these reasons do not explain all of the difference between actual net interchange and interchange modeled on CSCs.

Third, 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. For instance, the South-North interface is made up of the two 345 kV lines connecting the South and North zones, however, ERCOT has defined 19 CREs ("Closely Related Elements") which can also constrain flows from the South Zone to the North Zone. While ERCOT has the discretion to take CREs into account when managing interzonal congestion, they do not have the flexibility to do this efficiently. SPD always uses the CSC shift factors, although shift factors for CREs between the South Zone and North Zone may differ significantly from shift factors for the CSC. This leads to inefficient re-dispatch to manage constrained CREs.

Table 3 shows significant changes in the levels of net imports into each zone between 2003 and 2005. Imports to the Houston zone rose substantially from 2003 to 2004 and remained about the same from 2004 to 2005. The West Zone shifted from being a net exporter in 2003 to importing substantial quantities in 2004 and raising the level of imports in 2005. From 2003 to 2005, net exports increased from the South Zone as well as the combined area of the North and Northeast zones. In every case, the flows on CSCs were significantly less than the actual interchange in both years. In 2003, the Houston Zone showed the largest percentage difference, importing an average of 1,796 MW while SPD modeled an average CSC import of 565 MW.

Part of this difference occurred because the Houston Zone imported large quantities from the North Zone on four 345 kV transmission lines that were not managed by zonal balancing deployments in 2003. 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. To address this, ERCOT introduced the North to Houston CSC in 2004 to allow it to better manage interzonal congestion.[50] Table 3 indicates that SPD flows on CSCs into Houston more than doubled because of the addition of the North to Houston CSC. The North to West CSC was created in 2005 because ERCOT had frequently used OOME instructions to manage flows across the interface.

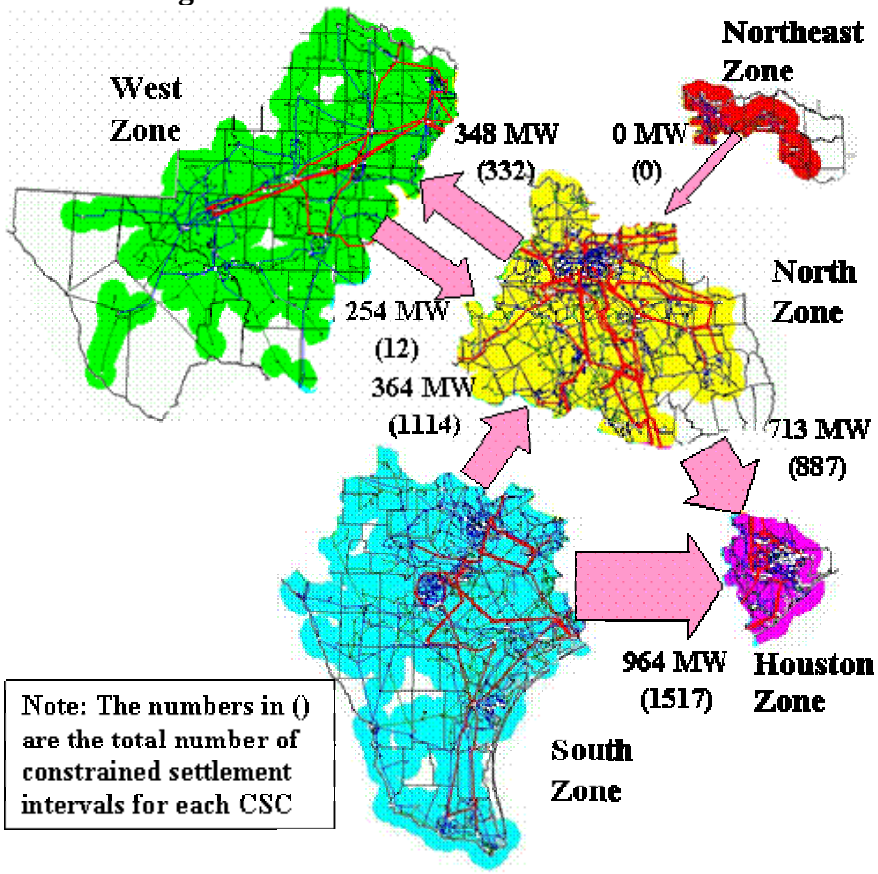
B. Interzonal Congestion

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 excludes most intervals, it is in these constrained intervals that the performance of the market is most critical.

Figure 58 shows the average SPD-calculated flows between the five ERCOT zones during constrained periods for the six 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.

[50] On an interim basis, the North to Houston CSC was managed using zonal OOME deployments from June to December 2003.

Figure 58: Average SPD-Modeled Flows on Commercially Significant Constraints During Transmission Constrained Intervals in 2005



Note: In the figure above, CSC flows are averaged taking the direction into account. For instance, if one hour has a North to West flow of 100 MW, and a second hour has a West to North flow of 200 MW, the average hourly flow would be 50 MW from the West to North. This treats the North to West flows in the first hour as negative for averaging purposes.

Figure 58 shows that inter-zonal congestion was most significant on the South to Houston CSC which exhibited SPD-calculated flows averaging 964 MW during 1,517 constrained intervals in 2005. Congestion was also significant on the South to North and North to Houston CSCs. The new North to West CSC experienced much more congestion than the West to North CSC which was congested for just 12 intervals during 2005. It is notable that the Northeast to North CSC was never constrained during 2005.

The lack of congestion on the Northeast to North CSC does not imply that having a Northeast zone that is distinct from the North zone has no impact on congestion management. When a CSC binds in the balancing energy market, SPD manages the constraint by redispatching generation in each zone. Due to the differences between zonal shift factors, each zone has a different impact on congestion. While the North and Northeast zones have similar shift factors,

decreasing generation in the North does more to relieve North to West CSC congestion than decreasing generation in the Northeast zone. Likewise, turning up generation in the Northeast zone provides more relief for South to North CSC congestion than increasing generation in the North zone.

1. Congestion Rights in 2005

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.

Market participants in ERCOT can hedge congestion charges 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 2005. Figure 59 shows the average number of TCRs and PCRs that were allocated for each of the CSCs in 2005, as well as the average SPD-modeled flows during the constrained intervals.

Figure 59: Transmission Rights vs. Real-Time SPD-Calculated Flows Constrained Intervals – 2005

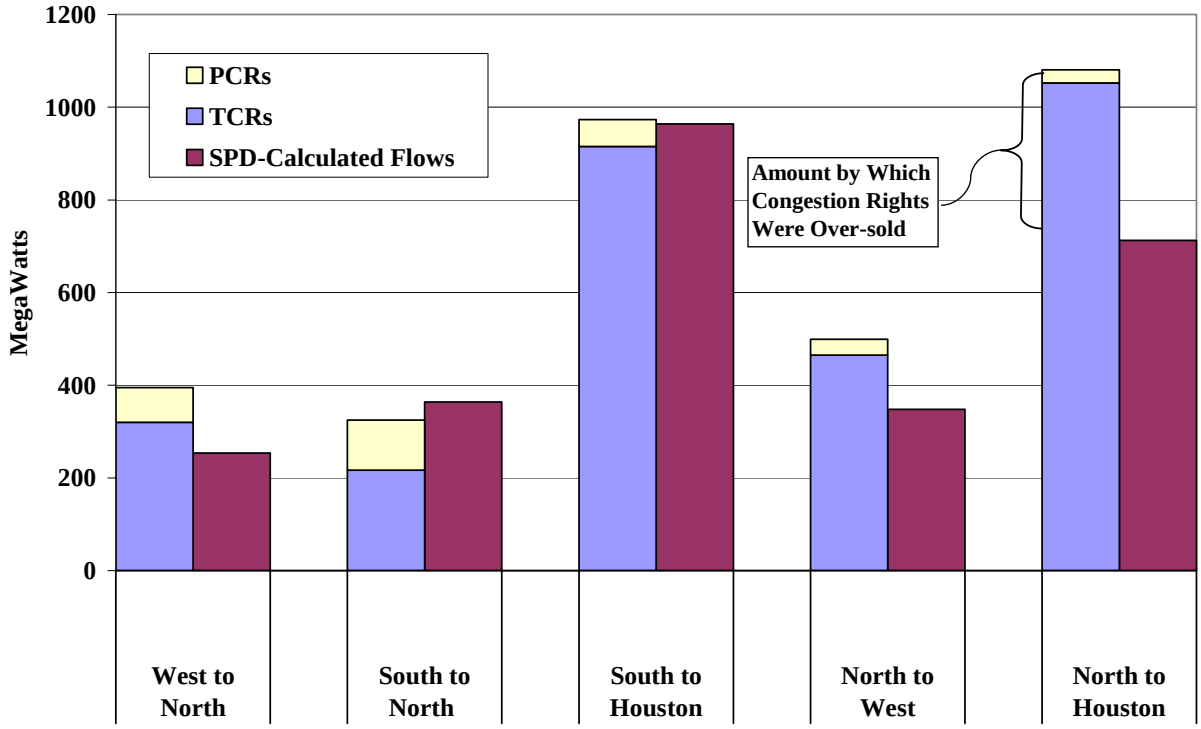


Figure 59 shows that total congestion rights (the sum of PCRs and TCRs) on the West to North, North to West, and North to Houston interfaces exceeded the average real-time SPD-calculated flows during constrained intervals. These results indicate that the congestion rights were oversold in relation to the SPD-calculated limits for some CSCs. For instance, congestion rights for the North to West CSC were oversold by an average of 151 MW.

The largest divergence between the SPD-calculated limits and the limits implied by the congestion rights was on the North to Houston CSC where 1,080 MW of congestion rights were allocated, but the average SPD-calculated flow during constrained intervals was 713 MW. Hence, the congestion rights that determine ERCOT's total obligation to make congestion payments exceeded the modeled flow over the CSC by an average of 367 MW.

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 ("BENA") charges.

To better understand the nature and causes of the shortfall implied by the results of Figure 59, we compare the SPD-calculated flows and congestion rights quantities for each of the constrained intervals by CSC.

2. Congestion on South to North CSC

The first CSC we analyze at the interval level is the South to North CSC. Figure 60 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 2005. Because only congested intervals are shown, some months will have significantly more observations than other months. Although some congestion occurred in every month, the three months from July to September accounted for 56 percent of all constrained intervals during 2005.

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 60 shows these changes in the congestion rights relative to the SPD-calculated flows, which fluctuate considerably in the congested intervals. In the figure, Total Congestion Rights include both TCRs and PCRs.

Figure 60: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to North – 2005

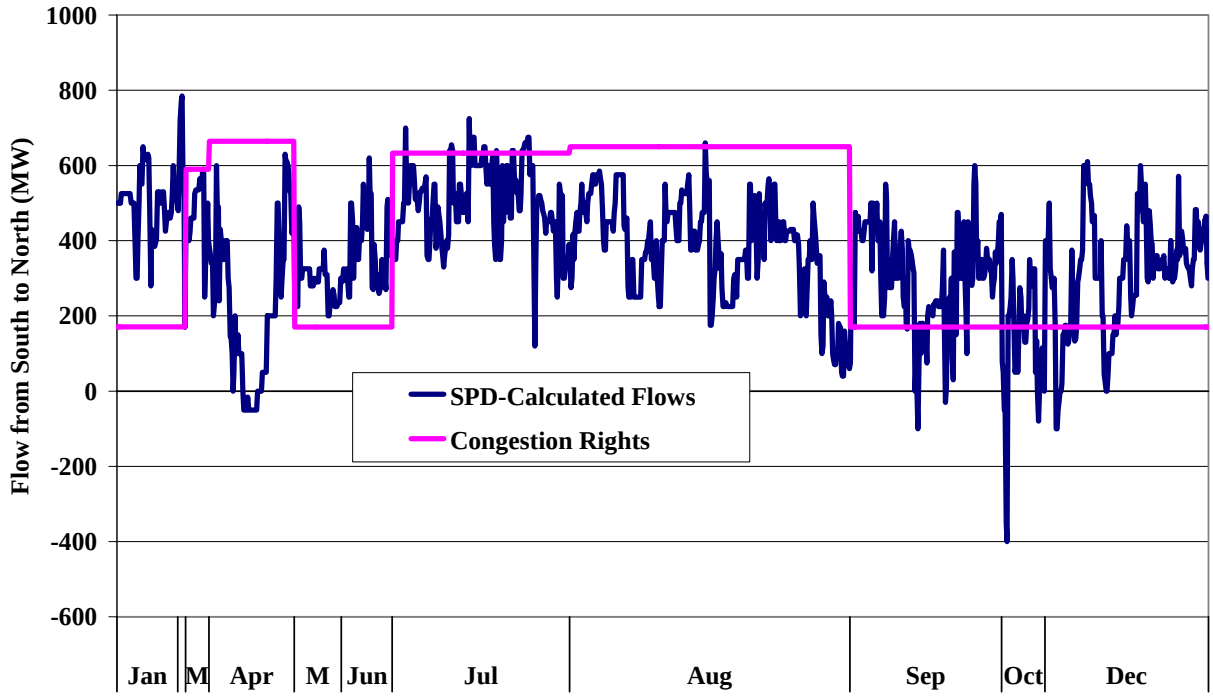


Figure 60 indicates that the quantity of outstanding congestion rights fluctuated considerably during 2005. From January to February, May to June, and September to December, fewer than 200 MW of rights were allocated for the South to North CSC, whereas in the other four months of 2005, approximately 600 MW of rights were allocated. This variation has to do with the complex nature of the South to North interface which results in it being constrained under a variety of circumstances.

Prior to each month, ERCOT estimates the transmission capability of the South to North interface based on transmission planning cases which use seasonal peak conditions. While two major lines make up the South to North interface, nearly 20 other transmission elements are defined as Closely Related Elements ("CREs"). Transmission constraints on the CREs can reduce the amount that can be transferred across the two major lines. Of the 12 monthly evaluations in 2005, there were seven distinct transmission elements that were found to be the most limiting. The pattern of flows can vary considerably, partly because of changes in the particular outages that are anticipated. Also, there is no guarantee that flows across the two main lines and all of the CREs will be in the same direction in every planning case. Indeed, there were five months in 2005 when ERCOT did not auction monthly TCRs because the estimated transfer

capability was negative. This occurred because the pattern of flows in the monthly planning cases resulted in power flowing from North to South while an element was constrained in the opposite direction. This highlights the issues that arise in the simplified zonal congestion management system. The nodal framework is better able to manage individual pieces of the transmission system, allowing more efficient utilization of the grid.

Figure 60 exhibits periods when SPD-calculated flows were both greater than and less than the quantity of congestion rights. Congestion rights nearly always exceeded SPD flows during the four months when approximately 600 MW of rights were allocated. Conversely, SPD flows usually exceeded the quantity of congestion rights during the eight months when less than 200 MW were allocated. The figure shows 29 constrained intervals when the SPD-calculated flows were *negative* during April, September, October, and December.

These very low SPD-calculated flows generally do not reflect the actual physical flows in real time, i.e., when the actual system conditions result in more flows over the South 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 South 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. Under extreme conditions, the operators must reduce the SPD limit into the negative range.

3. Congestion on Other CSCs

Like the South to North CSC, the other CSCs also exhibit significant differences between SPD-calculated flow limits and the quantity of outstanding congestion rights during congested intervals. The quantity of outstanding congestion rights can vary on a monthly basis according to projected transmission capability and the expected distribution of generation across the grid. SPD-calculated flow limits can change every 15 minutes according to system conditions and the mix of generation operating at the time. This section summarizes the results of our comparison of congestion rights to SPD flows during congested periods for the South to Houston CSC, the

North to Houston CSC, and the North to West CSC. The figures discussed in this section may be found in Appendix A.

The South to Houston CSC – With 1,517 constrained intervals, this interface experienced the most congestion of any CSC during 2005. The most congestion occurred in June, August, and September. In the months with significant congestion, SPD flows averaged between 950 and 1,050 MW. However, there was significant variation in the number of congestion rights allocated for this CSC by month, with as little as 761 MW in June when rights were under-sold substantially and 1,190 MW in August when they were over-sold.

The North to Houston CSC – This CSC was created in 2004 to manage congestion on a path into Houston that is usually able to physically transfer more than 2,000 MW. Prior to October 2005, ERCOT generally allocated between 1,100 and 1,300 MW of congestion rights for this CSC. In October 2005, however, the number of congestion rights was reduced below 500 MW and the frequency of transmission constraints rose dramatically. The primary reasons for this reduction in capability were the outages of the T.H. Wharton to Jewett 345 kV line, which is a component of the CSC, and the T.H. Wharton to O'Brien 345 kV line, which is a CRE.

The North to West CSC – This newly created CSC was congested primarily during the fall and winter months with approximately 75 percent of constrained intervals in January or February, or in the last three months of the year. Although the number of congestion rights allocated for this interface varied from 360 to 800 MW over the year, the SPD flows averaged just 341 MW during constrained intervals.

In conclusion, the quantity of available congestion rights can vary significantly due to outages and other changes in the topology of the transmission grid, and the distribution of generation and load. Furthermore, SPD-calculated flows can vary substantially and are frequently not close to the actual flows or limits for the CSC. Because transmission rights are generally sold based on the predicted 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 (without overselling) the available transmission capability.

One thing that generally improves the consistency between the quantity of congestion rights and SPD-calculated flows is consistency between the outages used to predict the availability of congestion rights and the actual outages that occur in real-time. Currently, ERCOT evaluates the availability of congestion rights approximately 35 days prior to the start of the month for which congestion rights are being auctioned. This leaves a significant amount of time for additional outages to be scheduled or occur involuntarily before real-time. Thus, ERCOT may be able to improve the consistency between the quantity of congestion rights and SPD-calculated flows during constrained periods by performing its evaluation less than 35 days before the beginning of the month for which the rights are being auctioned.

The New York ISO implemented a new settlement provision in 2004 to reduce inconsistencies between the quantity of congestion rights and the flows that actually occur.[51] Under the new provision, the ISO estimates the reduction in flows on congested interfaces due to transmission outages not considered in the congestion rights auction. The lost congestion rents associated with the reduced flows are charged to the transmission owner. This gives transmission owners better incentives to schedule maintenance during periods that have minimal market impact. A similar measure might also lead to more efficient maintenance scheduling if adopted in ERCOT.

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 40 percent of the transmission congestion rights sold in the annual auction. These rights are auctioned in December for the coming year. The remaining 60 percent of the transmission congestion rights are designated based on monthly updates of the summer study.[52] Since the monthly studies tend to more accurately reflect conditions that will prevail in

---

[51] In New York, financial congestion rights are called Transmission Congestion Contracts. These are sold in seasonal auctions and settle based on day-ahead market prices rather than real-time market prices.

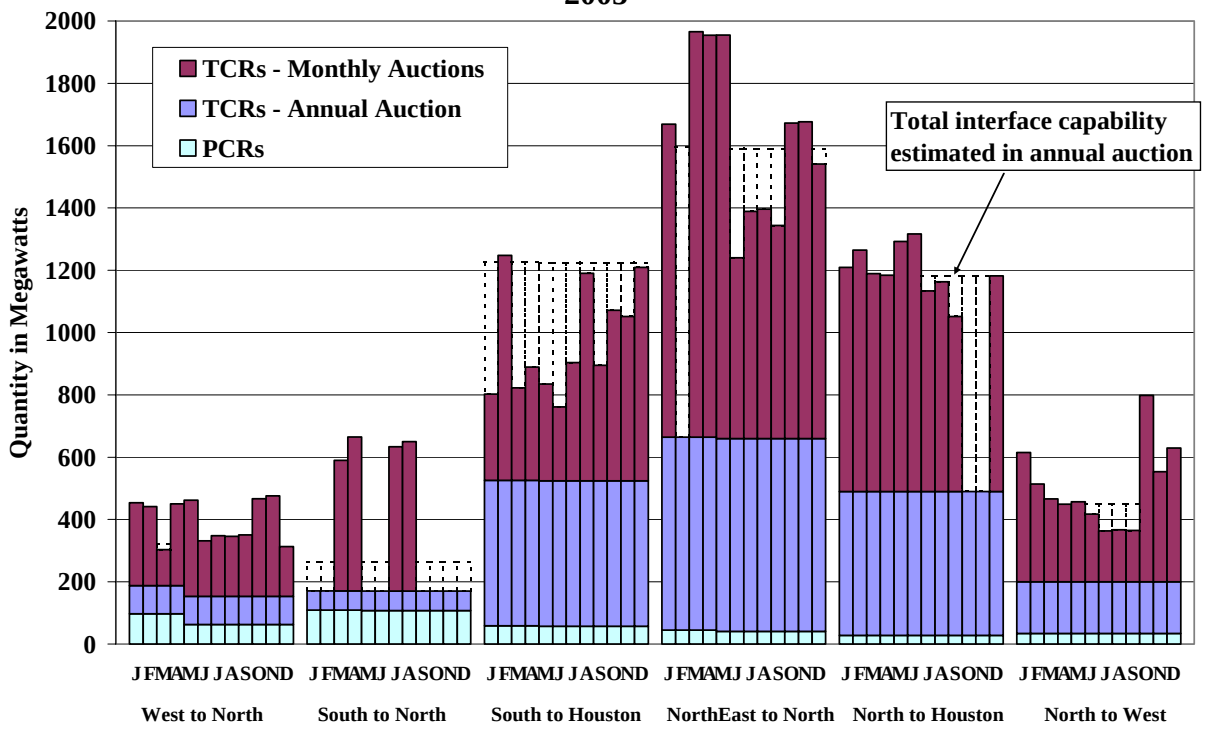
[52] Prior to 2005, 60 percent of estimated capability (after accounting for Pre-assigned Congestion Rights which are assigned to NOIEs) was sold in the annual auction. The remaining 40 percent was sold in the monthly auctions. This was changed because there were instances when the capability estimated before the monthly auction was more than 40 percent lower than the capability estimated before the annual auction. In these cases, no congestion rights could be sold in the monthly auction because no unsold capacity remained.

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 generation 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.

To examine how these processes have together determined the total quantity of rights sold over each interface, Figure 61 shows the quantity of each category of congestion rights for each month during 2005. The quantities of PCRs and annual TCRs are constant across months and were determined before the beginning of 2005, while monthly TCR quantities can be adjusted monthly.

Figure 61: Quantity of Congestion Rights Sold by Type
2005



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 61, 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.

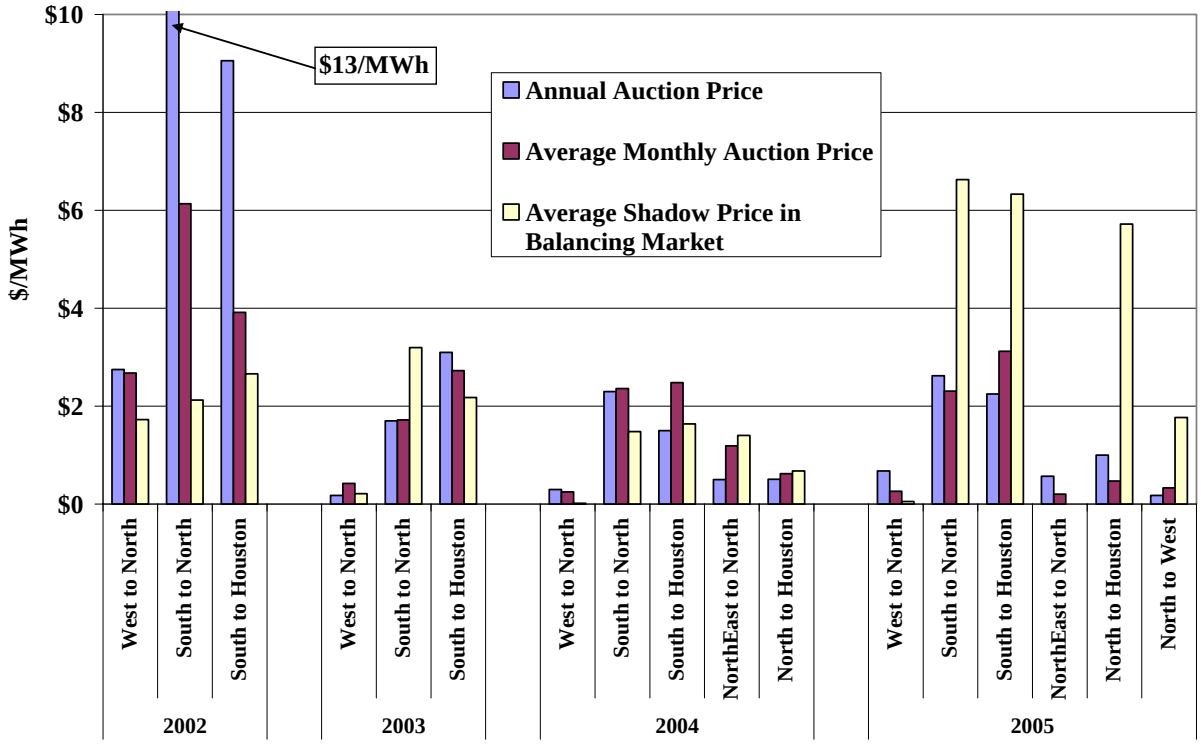
The South to North, North to Houston, and Northeast to North interfaces experienced the largest fluctuations in the estimates of transmission capacity from the annual auction to the monthly auction. In fact, South to North TCRs were not even auctioned during eight of the monthly auctions. There were also several instances when no congestion rights were available to be sold for the North to Houston CSC and the Northeast to North CSC in the monthly auctions. The divergence between annual and monthly estimates of transmission capacity on the other interfaces was smaller.

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 62 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 62 indicates that in 2002, the annual auction for the TCRs resulted in prices that substantially over-valued the congestion rights, particularly 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.

Figure 62: TCR Auction Prices versus Balancing Market Congestion Prices 2002 to 2005



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. It is noteworthy that the TCRs for the South to North and South to Houston interfaces settled at prices in 2004 that were closer to the previous year's value than in 2003. This indicates that participants have improved in their ability to forecast interzonal congestion and to value the TCRs, in part by observing historical outcomes. This improvement is likely facilitated by the simplified zonal representation of the ERCOT network embedded in the balancing energy market.

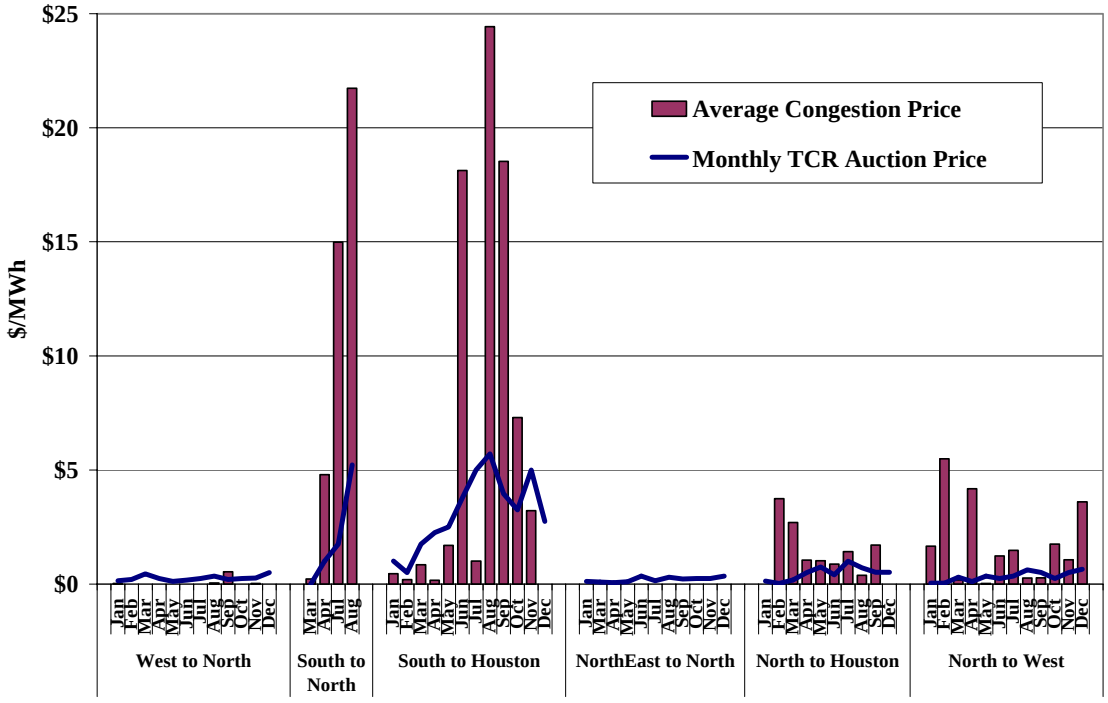
In 2004, TCR auction prices for the West to North, South to North, and South to Houston interfaces were similar to the previous year. Since congestion tends to be consistent across time, the auction prices for 2004 were reasonable predictors of real-time congestion. In 2004, there were two new products in the TCR auctions for the new CSCs. In both cases, the annual TCR price was below the monthly average TCR price, which was slightly below the average value of

congestion. This reflects cautiousness on the part of market participants when purchasing a TCR for a CSC that did not exist before 2004.

In 2005, market participants substantially under-estimated the value of congestion on the CSCs. The annual and monthly TCR prices in 2005 were generally in line with the TCR prices and the levels of balancing market shadow prices that prevailed in 2004. However, the actual volume and prices of congestion were substantially greater than in 2004, particularly on the South to Houston, South to North, and North to Houston CSCs. The North to West CSC was also substantially under-valued in the TCR auctions, although this is understandable given the lack of experience that market participants have with a newly created CSC.

Figure 63 compares monthly TCR auction prices with monthly average real-time CSC shadow prices from SPD for 2005. 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 CSC, the Northeast to North CSC, and the North-Houston CSC.[53]

Figure 63: Monthly TCR Auction Price and Average Congestion Value 2005



53 Notice that these missing months correspond to the missing monthly auction values in Figure 61.

Although congestion in the balancing market can be sporadic and inherently difficult to predict, the monthly TCR prices for the South to Houston interface exhibited patterns that were correlated to balancing market prices. After the South to Houston CSC began to experience very high levels of congestion during the summer, monthly TCR prices also rose indicating that market participants incorporated this information into their bidding behavior.

Overall, market participants did a poor job predicting fluctuations in congestion during 2005, particularly on the North to Houston and North to West interfaces. For both of these interfaces, there were several months when balancing market congestion spiked, far exceeding the TCR prices in those months. However, based on the TCR prices, there is little sign that market participants expected an increase in congestion in those months relative to other months.

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 $33,000 (600 MWh * $55/MWh) while suppliers in the West Zone will receive $24,000 (600 MWh * $40/MWh). The net result is that ERCOT collects $9,000 in congestion rent ($33,000 – $24,000) and uses it to fund payments to holders of TCRs.

[54] 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 64 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 64: TCR Auction Revenues, Credit Payments, and Congestion Rent[55] 2002 to 2005

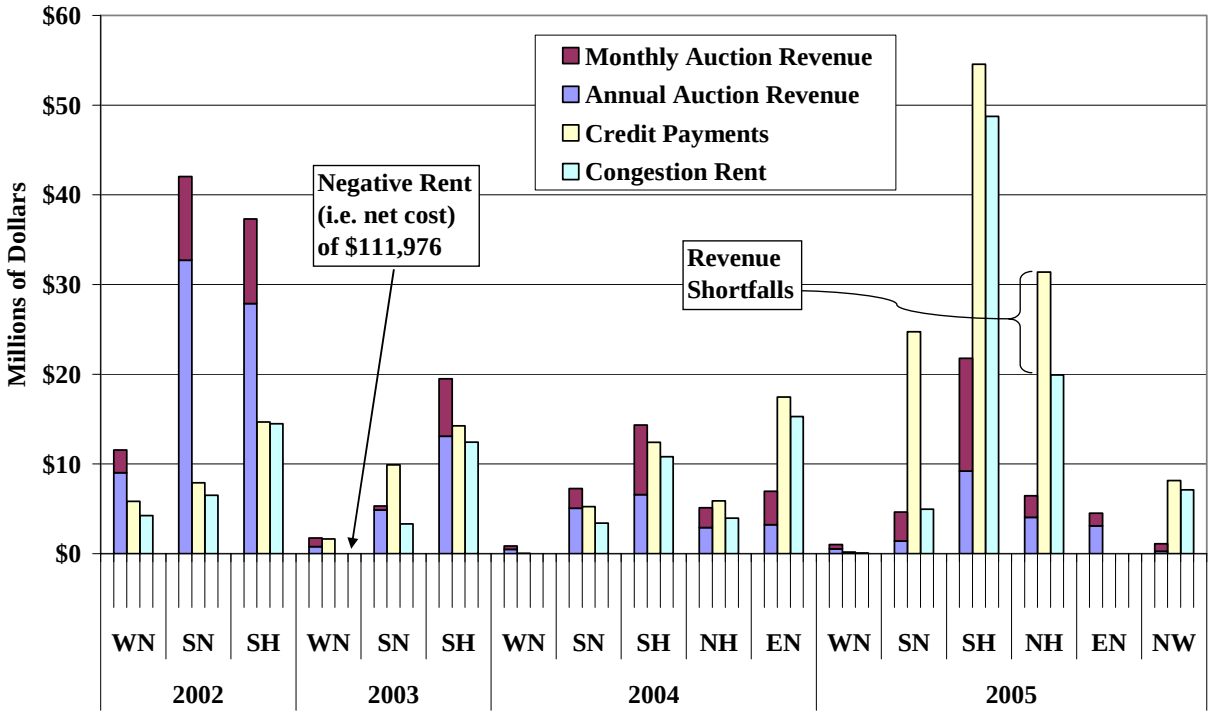


Figure 64 shows that in 2002, the total auction revenues were far greater than credit payments to TCR holders. This is the result of the auction prices being much greater than the average shadow prices that occurred in the balancing energy market (as was shown in Figure 63 above). The

[54] This explanation is simplified for the purposes of illustration. However, congestion rents would also depend on the net imports into and net exports from the other three zones as well as the zonal prices. Furthermore, the net exports from the West Zone do not necessarily match the net imports into the North Zone in real-time operation.

[55] The source for congestion rents is the ERCOT TCR Program Report. However, this source incorporates an additional term based on the revenue impact of using generation-weighted shift factors for loads instead of the load-weighted shift factor.

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.

In 2004, the auction revenues were consistent with credit payments for the three CSC that existed in 2003. This appeared to be due to market participant basing their valuations of the TCRs on their value in prior years. The auction revenues for the North to Houston CSC, which was added for the first time in 2004, were quite close to credit payments. However, market participants substantially under-valued congestion on the Northeast to North interface, which was also new in 2004.

In 2005, the auction revenues were greatly exceeded by credit payments for the four interfaces with significant congestion. This was because the TCR market under-estimated the volume of congestion that would occur in the balancing market. TCR prices were generally consistent between 2004 and 2005, suggesting that market participants based their expectations on the levels of congestion that occurred in 2004. Since interzonal congestion in balancing market was far greater in 2005 than in previous years, payments to TCR holders exceeded TCR auction revenues by a significant margin.

Figure 64 also shows that payments to TCR holders have consistently exceeded the congestion rents that have been collected from the balancing market since the creation of the TCR market. The difference was relatively modest in 2002 when congestion rents covered 93 percent of payments to TCR holders and in 2004 when they covered 81 percent. However, in 2003 and 2005, congestion rents covered only 61 percent and 68 percent of payments to TCR holders. In 2005, the gap between congestion rents and payments to TCR holders resulted in a $38 million revenue shortfall. When congestion rents fall significantly below payments to TCR holders, it implies that the SPD-calculated flows across constrained interfaces have been systematically lower than the amount of TCRs sold for the interfaces.

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.[56] These shortfalls are included in the Balancing Energy Neutrality Adjustment charge and assessed to load ERCOT-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 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 manual 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.

---

[56] 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.

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 actual output resulting from 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).

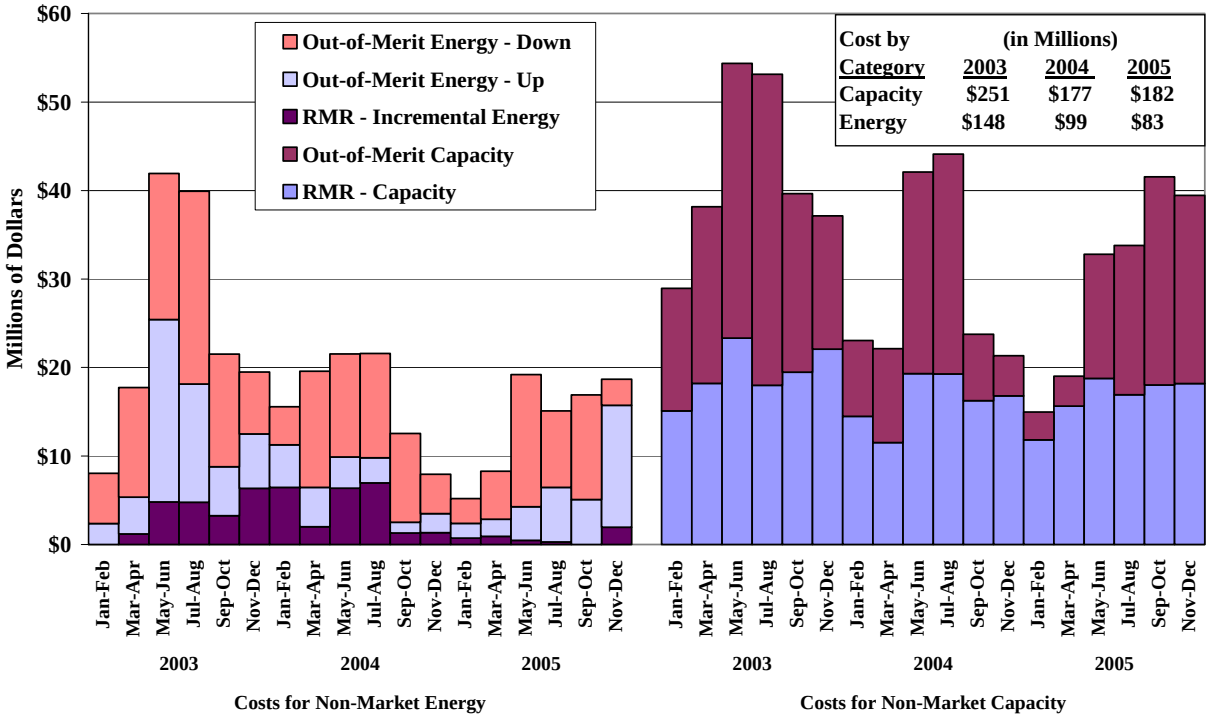
For OOME down, the Protocols establish an avoided cost level based on generation type that determines the OOME down payment obligation to the participant. If a unit with an avoided cost under the Protocols of $15 per MWh receives an OOME down instruction when the balancing energy price is $35 per MWh, then ERCOT will make an OOME down payment of $20 per MWh.

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. Since 2002, several units have acquired RMR status while several others have been taken off. Initially, all of the RMR units were located in the South and West zones, but in 2004, units at the Eagle Mountain plant in the North zone were added to RMR status. In 2005, one of the PH Robinson units in Houston was given RMR status, while all units in the West zone were taken off RMR

status. Units contracted to provide RMR service to ERCOT are compensated for start-up costs, energy costs, and are also paid a standby fee. 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 to the minimum dispatch level, and (b) incremental energy costs, which are the costs associated with output above the minimum dispatch level. Figure 65 shows each of the five categories of uplift costs from 2003 to 2005.

Figure 65: Expenses for Out-of-Merit Capacity and Energy 2003 to 2005



The left side of Figure 65 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.

The results in Figure 65 show that overall uplift costs for RMR units, OOME units, and OOMC units were relatively consistent between 2004 and 2005, following a steep decline of $399 million to $276 million from 2003 to 2004, a decrease of 31 percent. From 2003 to 2005, there were substantial percentage reductions in uplift payments to OOMC units, OOME-up units, and OOME-down units, all of which were reduced by nearly 40 percent over the period. Out-of-

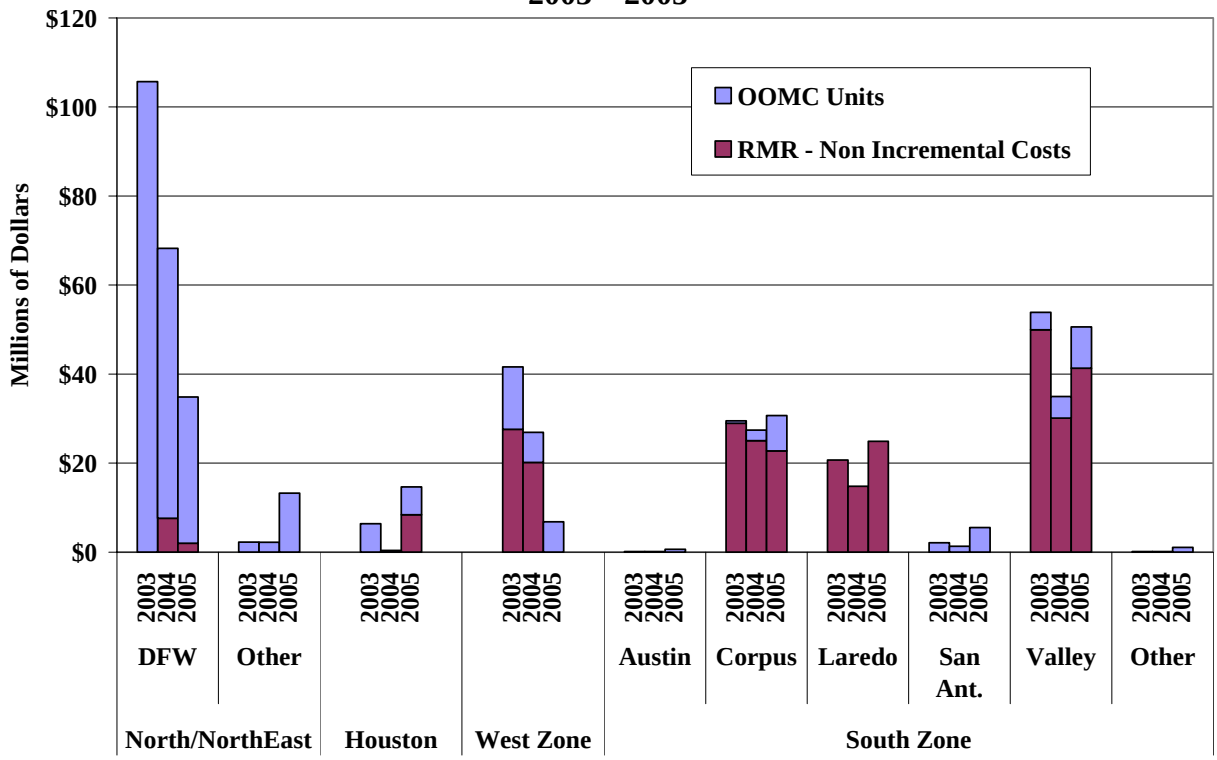
merit costs have been greater during the summer when higher loads have increased the need for ERCOT operators to take out-of-merit actions to manage local congestion and reliability needs. Conversely, RMR costs have been less variable during the year because RMR payments are primarily designed to recover fixed costs, which are constant throughout the year.

In 2005, the seasonal variation in uplift costs was different from the two previous years. While uplift costs were very low during the spring and summer of 2005, uplift costs actually rose considerably during the final four months of 2005. The figure above indicates that uplift cost totaled $116 million in the last four months of 2005, up from $66 million during the same period in 2004. However, total uplift payments from January to August 2005 were 29 percent lower than the same period in 2004.

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 between 2003 and 2005. The first of these analyses focuses on the payments made for commitment of capacity, which include OOMC payments and net RMR payments excluding the portion for incremental energy dispatch. Figure 66 shows these payments by location.

Commitment-related uplift costs decreased dramatically from 2003 to 2005 in the Dallas/Ft. Worth ("DFW") area and in the West zone. In DFW, the reduction was due to less frequent OOMC commitments, whereas uplift was reduced in the West zone by the elimination of RMR status for units there. Most regions in the South zone have had relatively consistent uplift costs over time, primarily because RMR units are the biggest source of uplift costs there. Prior to 2005, Houston and the North zone outside DFW were the source of very little commitment-related uplift, but in 2005, these areas accounted for approximately $30 million.

Figure 66: Expenses for OOMC and RMR by Region
2003 – 2005



ERCOT began to use dynamic (rather than static) ratings for major transmission elements in March 2005, and this has likely helped reduce the need for out-of-merit commitment and dispatch. The thermal capability of transmission elements decreases as ambient temperatures increase. Without real-time information on the capacity of individual elements, the system operator must make conservative assumptions in order to keep the system reliable. The dynamic ratings allow ERCOT to maximize the transfer capability of the transmission system, while maintaining reliability. This has likely resulted in significant gains in efficiency and cost savings for consumers.

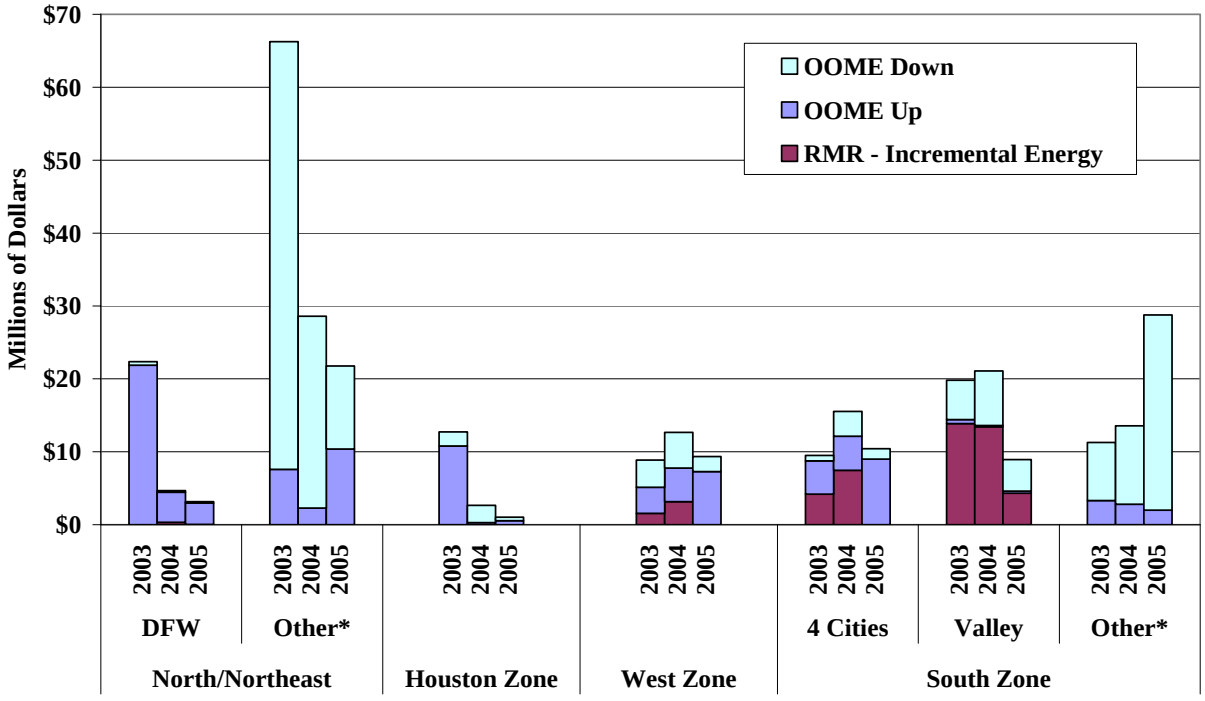
Significant transmission upgrades in the Dallas/Fort Worth area have led to significant reductions in OOMC costs associated with that area after 2003. Changes to the compensation formulas for OOMC units in February 2004 also contributed to reductions in OOMC costs. According to the ERCOT Protocols, ERCOT pays OOMC units for (a) starting-up and (b) staying on-line. Previously, payment formulas for starting-up and staying on-line were not dependent on the balancing energy price, which caused problems for two reasons. First, there

was less likelihood that the sum of the start-up payment, operating payment, and revenue from the balancing energy market would be sufficient for a unit to recover its costs.

Second, units would receive the same uplift payment regardless of whether the balancing energy market revenue at the prevailing price was compensatory. This created a disincentive for QSEs to voluntary commit resources that were frequently needed for local reliability. It was often more profitable to wait for the resources to be committed through the OOMC process. We believe the current formulas have reduced this incentive problem by making it less profitable to have ERCOT commit resources through the OOMC process when prices are expected to be high enough to cover the resources' commitment costs.[57] This change has likely contributed to the lower OOMC commitment costs in 2004 and 2005.

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

Figure 67: Expenses for OOME by Region
2003 to 2005



* The "Other" category includes portfolio deployments where the operator does not specify a single resource.

[57] The new compensation formulas were defined by PRRs 371 and 450.

well as payments to RMR resources for incremental energy above minimum generation.[58] Figure 67 shows annual uplift costs for units providing OOME by region and by zone.

The figure summarizes the overall costs of out-of-merit energy (including the portion from RMR units). These have varied over the past three years due to several factors. First, in 2004, ERCOT created two new CSCs in order to reduce the need for OOME dispatch. The Northeast to North CSC was created because ERCOT was frequently giving OOME-down deployments to generators in the northeast region and OOME-up deployments to generators in and around DFW. After the CSC was created, suppliers have had an incentive to resolve congestion between the Northeast and North zones because the costs of congestion are directly assigned to the generators that cause it. The North to Houston CSC was created to give participants in the balancing market the incentive to resolve the congestion on the 345 kV lines directly connecting the North zone to Houston.

Second, changes in the quantity of fleet-level OOME deployments account for some of the changes in uplift from 2003 to 2005. Several figures in Section II showed that substantial amounts of on-line capacity are not offered to the balancing market, even after accounting for what is set aside to satisfy ancillary services obligations. When ERCOT runs short of balancing energy offers, and it is not possible to deploy non-spinning reserves in the necessary timeframe, fleet-level OOME deployments are used by ERCOT to maintain the balance of supply and demand and/or manage congestion on CSCs. Thus, fleet-level OOME deployments are not made to manage local congestion.

Third, we also attribute some of the reduction in local congestion costs in 2004 and 2005 to the suspension of the "Market Solution" method used to solve local transmission constraints. Prior to July 18, 2003, a Market Solution would exist whenever three or more unaffiliated suppliers were capable of relieving a local constraint. When a Market Solution existed, SPD would select the most cost effective resource(s) based on the shift factors and offer premiums for each resource. Incremented resources were paid the energy clearing price plus their offer premium. Likewise during this period, ERCOT paid decremented resources their offer premium as compensation for the lost opportunity of producing at the market price. Market Solutions

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

accounted for approximately $22.8 million in uplift payments from January to July 2003 to relieve relatively small amounts of congestion in real-time. ERCOT discontinued the Market Solution process because, the outcomes often were non-competitive and ERCOT frequently dispatched resources with offer premiums approaching $1,000/MWh.

E. Conclusions: Interzonal and Intrazonal Congestion

Consistent with our conclusions from previous reports, the results in this section highlight significant opportunities for improvements in the operation of the ERCOT markets.[59] These results indicate that, although inter-zonal congestion rose considerably in 2005, the majority of the congestion costs are still associated with intrazonal congestion. ERCOT has taken steps which have helped reduce intrazonal congestion substantially since 2003. However, the process of intrazonal congestion management 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, this section highlights significant issues related to the zonal assumptions used in the ERCOT market. These assumptions and the operation of the current markets in Texas were evaluated in greater detail in the Market Operations Report issued in 2004, which identified some significant issues related to congestion management processes in ERCOT [60] In addition, the results in this section of the report continue to indicate 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, leading to $38 million in uplift costs in 2005, a substantial increase from previous years; and
- These inconsistencies can result in under-utilized transmission capability and difficulties in defining transmission rights whose obligations can be fully satisfied.

The long-run remedy for both the interzonal and intrazonal issues identified in this report will be the implementation of nodal markets. The nodal markets currently being designed for

---

59 See 2003 SOM Report, Assessment of Operations, and 2004 SOM Report

60 See "2004 Assessment of the Operation of the ERCOT Wholesale Electricity Markets", Potomac Economics, November 2004.

implementation in 2009 will provide transparent prices for both generators and loads that would fully reflect all transmission constraints on the ERCOT network.

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 2005. We examine market structure using a pivotal supplier analysis, which indicates that the two largest suppliers in ERCOT were frequently pivotal in 2005. This analysis also shows that the frequency with which a supplier was pivotal increased with the level of demand. To evaluate participant conduct, we estimate measures of physical and economic withholding. We examine withholding patterns relative to the level of demand and the size of each supplier's portfolio. Based on theses analyses, we find patterns that are suggestive of economic withholding that raise competitive concerns.

A. Structural Market Power Indicators

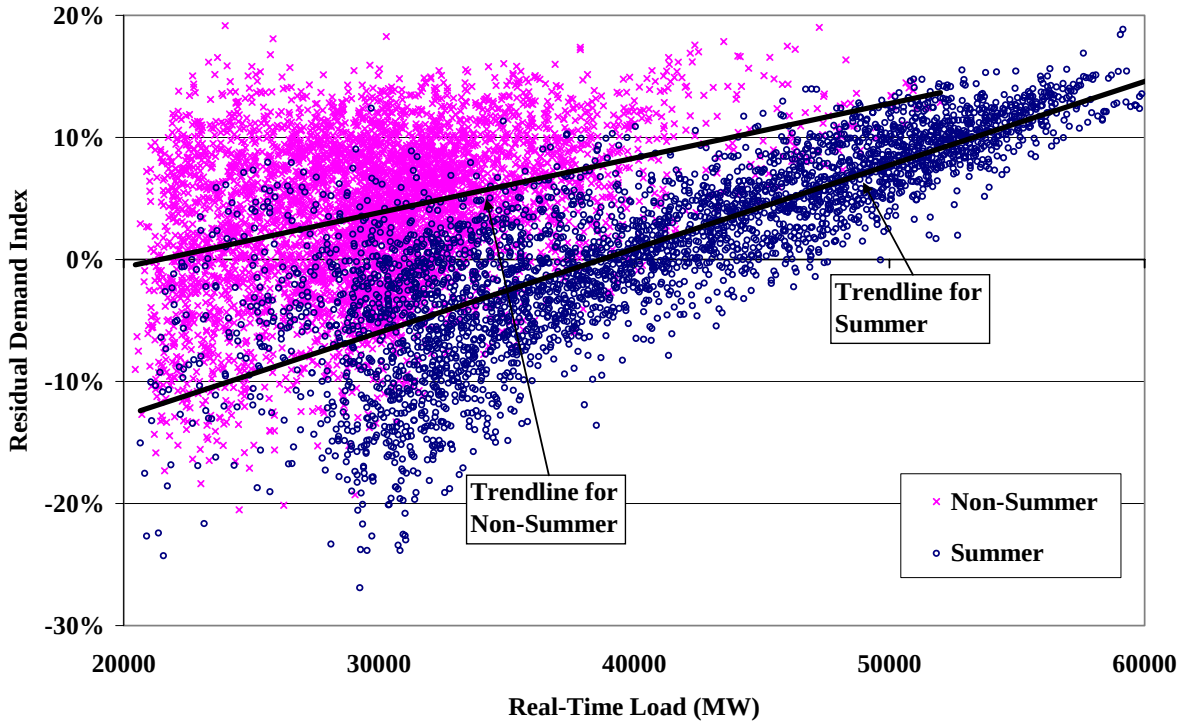
We analyze market structure using the Residual Demand Index ("RDI"), a statistic that measures the percentage of load that could not be satisfied without the resources of the largest supplier. When the RDI is greater than zero, the largest supplier is pivotal (i.e. its resources are needed to satisfy 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 68 shows the RDI relative to load on an hourly basis in 2005. The data is divided into two groups: (i) hours during the summer months (from May to September) are shown using darker points, while (ii) hours during other months are shown using lighter points. 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 QSEs would have to increase as load increases to keep the RDI from increasing. This analysis is done at the QSE level because the largest suppliers that determine the RDI values

shown below own a large majority of the resources they are scheduling or offering. It is possible that they also control the remaining capacity through bilateral arrangements, although we do not know whether this is the case. To the extent that the resources scheduled by the largest QSEs are not controlled or providing revenue to the QSE, the RDIs will tend to be slightly overstated.

Figure 68: Residual Demand Index
2005



The figure shows that the RDI for the summer (i.e. May to September) was usually positive in hours when load exceeded 40,000 MW. During the summer, the RDI was greater than zero in approximately 57 percent of hours. During the non-summer period, the RDI was generally positive under all load conditions. The RDI was typically positive at lower load levels during the spring and fall due to the large number of generation planned outages and less commitment. Hence, although the load was lower outside the summer, our analysis shows that a QSE was pivotal in almost 75 percent of hours during that period. In addition to being higher on the page, the non-summer trend line exhibits a flatter slope than the trend line for the summer period. The flatter slope of the non-summer trend line indicates a weaker relationship between the RDI and demand level in the fall. The figure shows RDI values that were generally higher in 2005 than in 2004 primarily because of the reduction in excess on-line and quick start capacity.

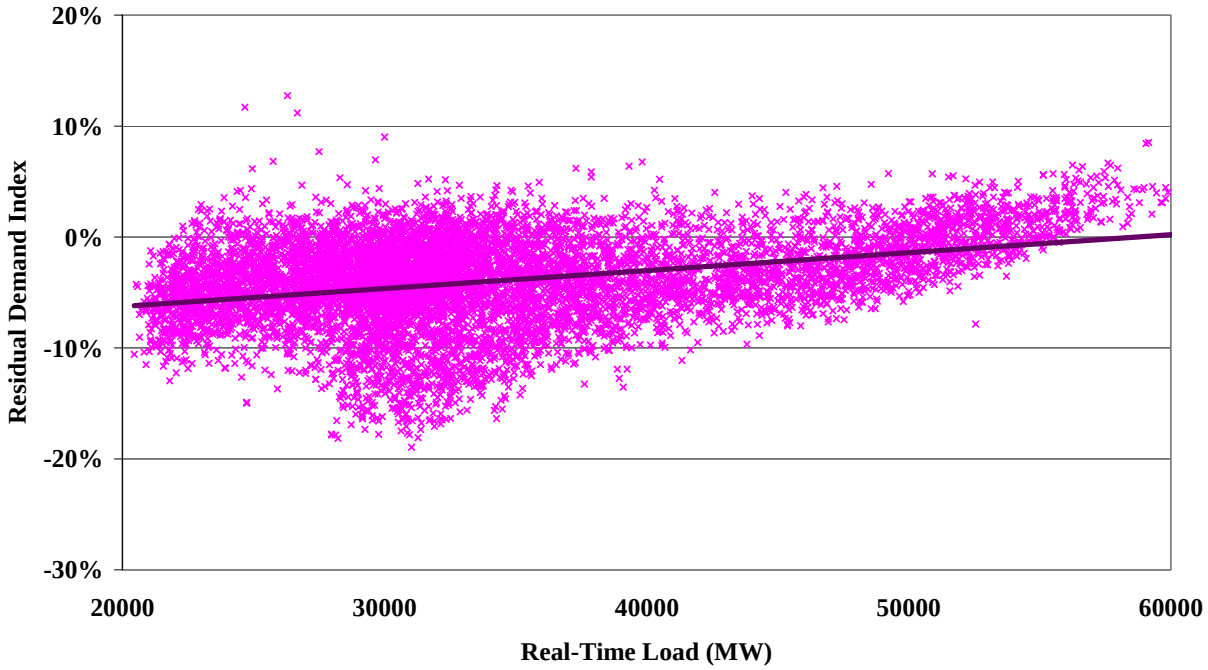
It is important to recognize that inferences regarding market power cannot be made solely from this data. Retail load obligations can affect the extent of market power for large suppliers, since such obligations cause them to be much smaller net sellers into the wholesale market than the analysis above would indicate. Bilateral contract obligations can also affect a supplier's potential market power. 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 substantial long-term sales contracts. The RDI measure shown in the previous figure 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 energy market may be higher than indicated by the standard 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 according to the standard RDI shown above. To account for this, we developed RDI statistics for the balancing energy market. Figure 69 shows the RDI in the balancing energy market relative to the actual load level.

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[61] owned by other suppliers. Figure 69 limits the other supplier's capacity to the capacity 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 shortage of offers in the balancing energy market.

[61] For the purposes of this analysis, "quick-start" includes off-line simple cycle gas turbines that are flagged as on-line in the resource plan with a planned generation level of 0 MW that ERCOT has identified as capable of starting-up and reaching full output after receiving a deployment instruction from the balancing energy market.

Figure 69: Balancing Energy Market Residual Demand Index vs. Actual Load 2005



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 69 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 QSEs offering 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 standard RDI is negative). 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. We discuss this subject in greater detail in Section II of this report.

B. Evaluation of Supplier Conduct

The previous sub-section presented a structural analysis that supports inferences about potential market power. In this section we evaluate actual participant conduct to assess whether market participants have attempted 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 also 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 foregone sales of its withheld capacity.

1. Evaluation of Potential Physical Withholding

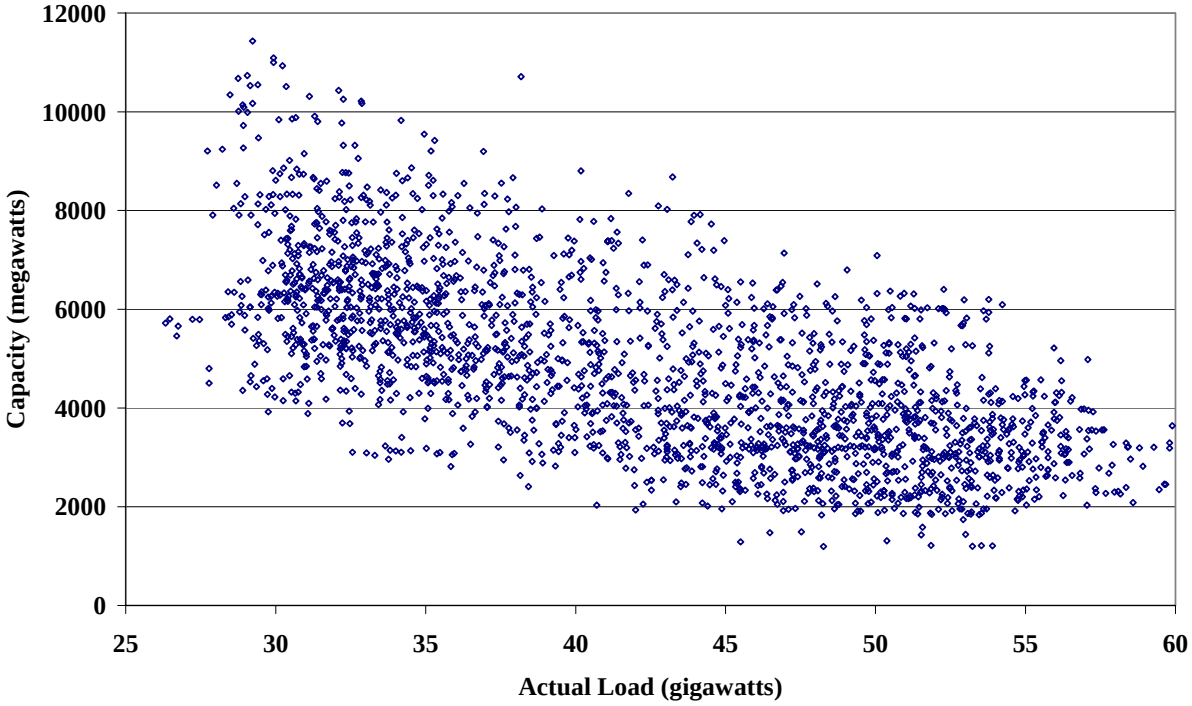
Physical withholding occurs when a participant makes resources unavailable for dispatch that are otherwise physically capable of providing energy and that are economic at prevailing market prices. This can be done by derating a unit or designating it as a forced outage. In any electricity market, deratings and forced outages are unavoidable. The goal of the analysis in this section is to differentiate 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 68 and Figure 69 indicate that the potential for market power abuses rises as load rises and RDI values become more positive. Hence, if physical withholding is a problem in ERCOT, we would expect to see increased deratings and forced outages at the highest load levels. Conversely, 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 70 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 unlikely to constitute physical withholding given the cost 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 70: Short-Term Deratings and Forced Outages vs. Actual Load
June to August, 2005



As the figure shows, short-term deratings and outages varied between 1 GW and 11 GW. Since the figure includes data from only three summer months, the lower load levels generally represent shoulder hours during weekends and nighttime. It is common for several QSEs to submit resource plans for some units with status "unavailable" during shoulder hours and status "available" during the daytime. This causes the data to show an inverse relationship between deratings and outages and real-time demand levels. At demand levels above 57 GW, the sum of deratings and outages were always 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 70 suggest 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 71 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.[62] The large supplier category includes the three 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).

Figure 71: Short-Term Deratings by Load Level and Participant Size
June to August, 2005

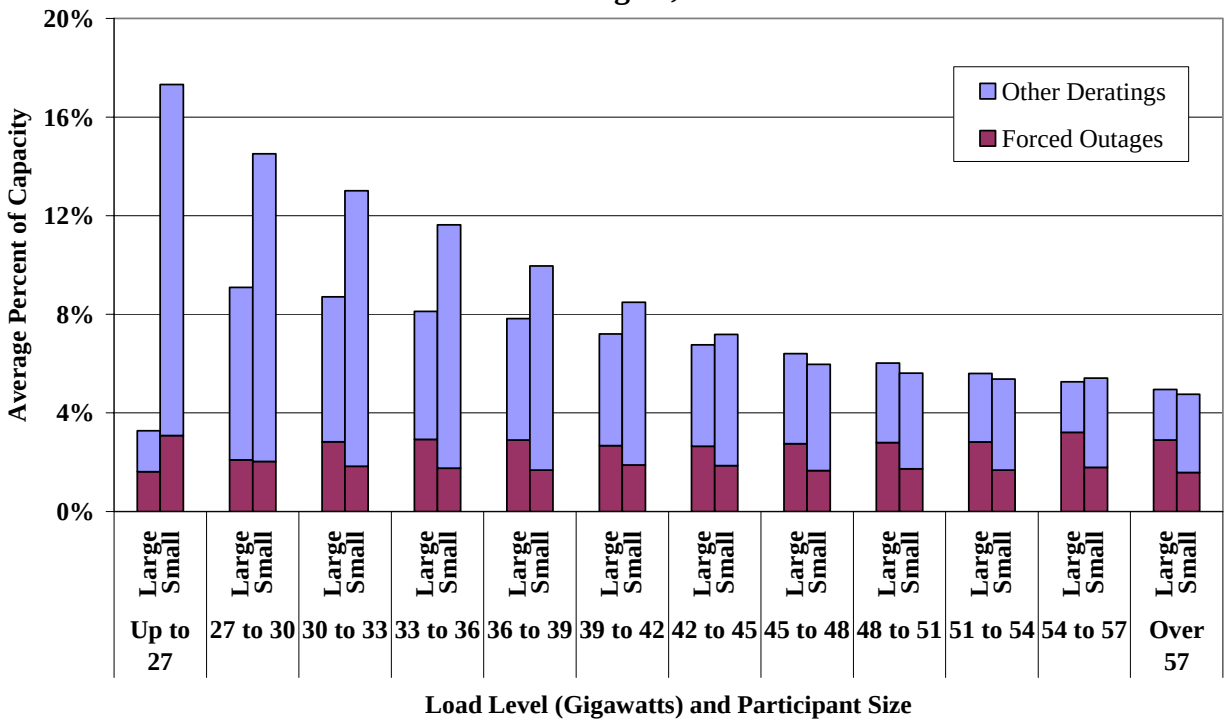


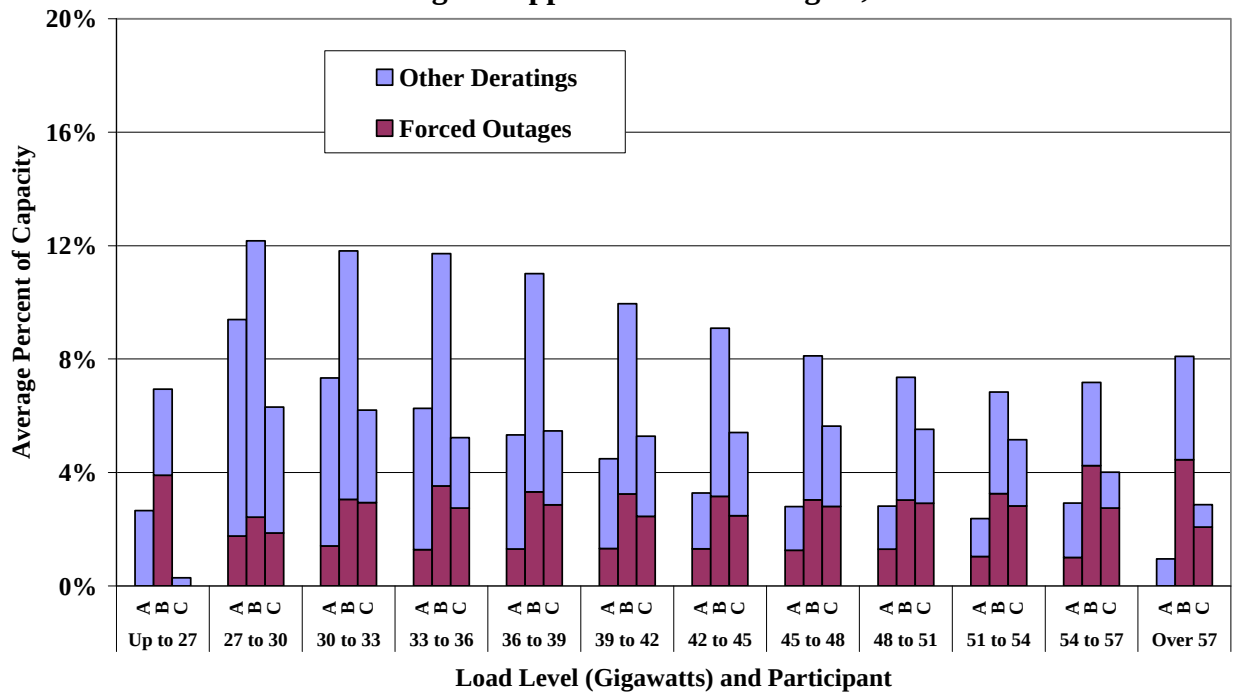
Figure 71 suggests that as electricity demand increases, both large and small market participants tend to make more capacity available to the market. For large suppliers, the short-term derating or forced outage rates decreased from approximately 8 percent at low demand levels to about 4 to 5 percent at load levels above 54 GW. For small suppliers, the derating rates decreased from 12 to 17 percent at load levels below 36 GW to less than 5 percent at load levels above 54 GW. As discussed above, the higher "other deratings" during lower load periods reflects a practice by

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

some smaller QSEs of designating some of their resources unavailable during weekends and nighttime periods. Figure 71 also shows a distinction between forced outages and other deratings and indicates 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.

At all load levels, large suppliers have deratings rates that are less than or equal to those of small suppliers. Furthermore, large suppliers' deratings and outages decline as load levels increase. Given that the market is most vulnerable to market power at the highest load levels, these derating patterns do not provide evidence of physical withholding by the large suppliers. However, these results cannot exclude limited instances of withholding by either large or small suppliers. The deratings of the large suppliers are examined in greater detail in Figure 72.

Figure 72: Short-Term Deratings by Load Level and Participant
Three Largest Suppliers – June to August, 2005



The three largest suppliers differ from one another in the fraction of their portfolios that are derated or on outage in the short-term. Company A has the lowest level of deratings, followed by Company C, while Company B has a substantially higher rate than the other two suppliers. By itself, this does not suggest physical withholding on the part of Company B since the levels of

deratings and outages vary according to the characteristics of the suppliers' portfolio. Older units and certain technology types are more likely to encounter operational issues that make it necessary to derate the unit. For instance, high ambient temperatures tend to reduce the capability of combustion turbines more than steam turbines. The three suppliers are similar in the sense that the short-term deratings and outages of all three exhibit a downward trend as load increases. This trend is a positive indication that the largest suppliers have not engaged in physical withholding. However, based on the figures above, we cannot definitively conclude whether or not market participants have engaged in physical withholding. Firm conclusions in this regard would require a more detailed investigation of the suppliers' outages and deratings.

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. The output gap excludes capacity that is necessary for the QSE to fulfill its ancillary services obligations. 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.[63]

[63] 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 input at minimum generation levels for ERCOT generating units. This analysis also assumes $4 per MWh

As was the case for 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 73 shows the relationship between the output gap from committed resources and real-time load for all hours during 2005.

Figure 73: Output Gap from Committed Resources vs. Actual Load 2005

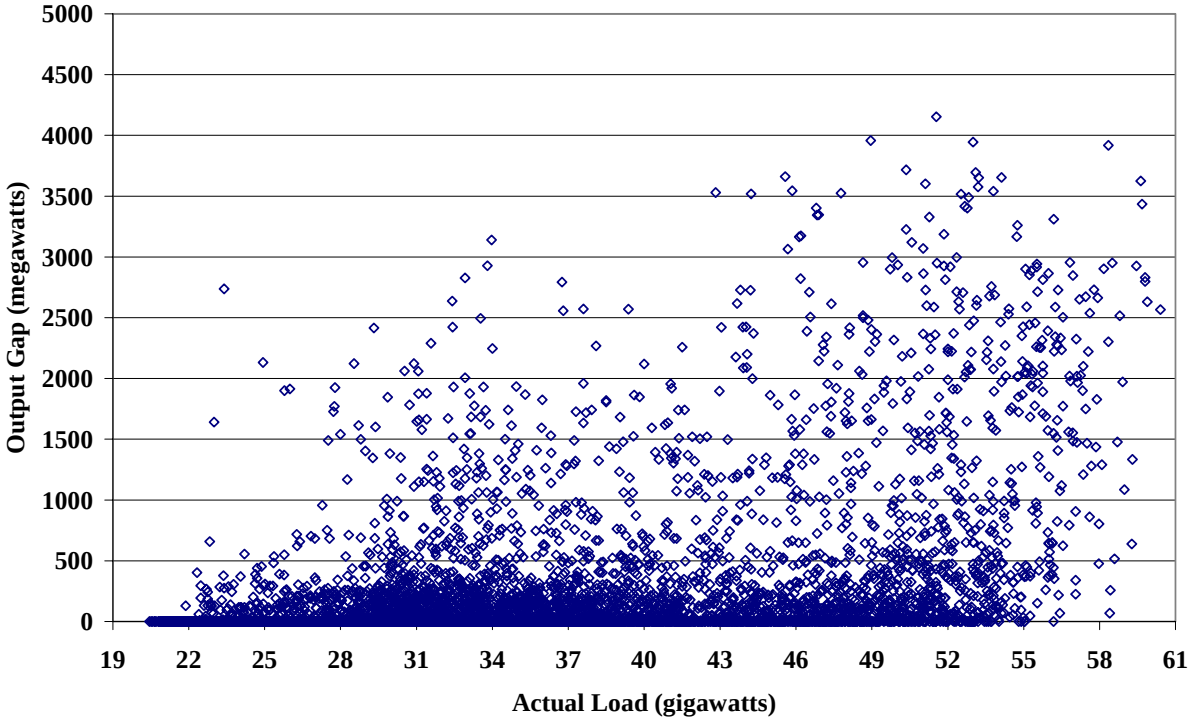


Figure 73 shows that the output gap from committed resources ranged from zero in most hours to a maximum of over 4,000 MW during 2005. This figure shows that the total output gap in the market tends to rise with real-time demand. This is not surprising given that clearing prices tend to be higher at higher load levels and we routinely observe thousands of megawatts of un-offered capacity. Many of the high output gap values occurred during transitory price spikes under a wide range of demand levels that 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 prevent resources from

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.

responding instantaneously to an unpredicted price spike. Even quick-start resources are sometimes unable to come on-line quickly enough to an unforeseen transitory price spike. The next analysis further examines the output gap results by size of supplier and load level.

Figure 74 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 three largest suppliers in ERCOT, whereas the small supplier category includes the remaining suppliers that each controls 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 74: Output Gap by Load Level and Participant Size 2005

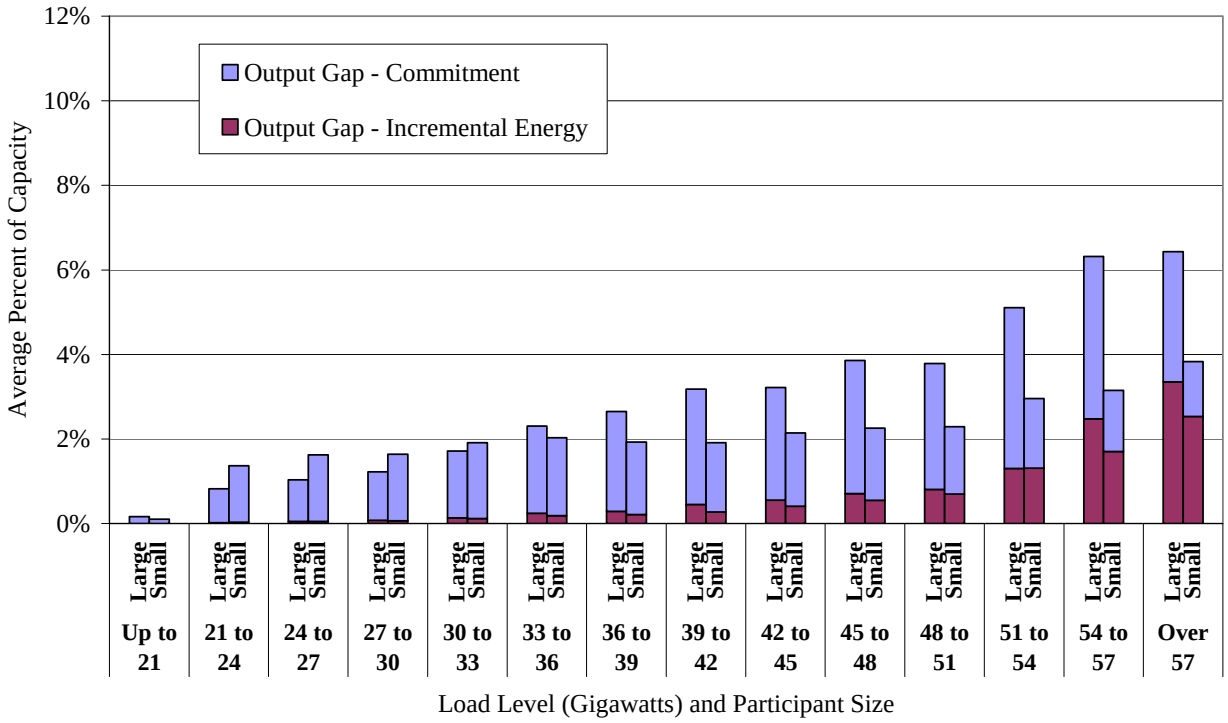
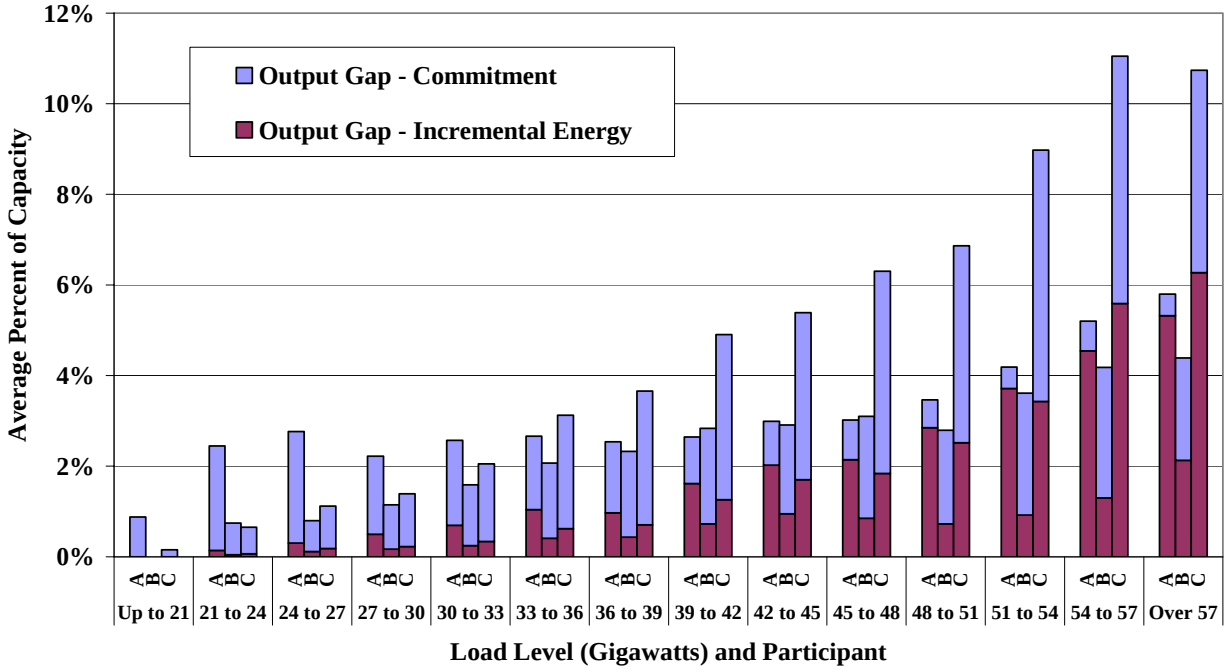


Figure 74 shows that the output gap quantities of large and small suppliers were comparable at lower load levels. However, large suppliers had substantially higher output gaps at higher load levels, particularly when load was greater than 51 GW. The greater output gaps for large suppliers were driven primarily by the failure to commit economic resources rather than the failure to fully dispatch on-line resources. Generally, market power is more likely to exist when

demand is relatively high, so the output gap results in the figure above raise some concerns about the potential for strategic withholding.

Large suppliers' output gap increased from close to zero at low demand levels to over six percent at the highest demand levels. For small suppliers, the output gap increased in a similar pattern from close to zero at low demand to almost four percent at the highest load levels. Due to the comparatively high output gap values for large suppliers, we performed two additional analyses to more closely examine the output gap patterns of large suppliers. The following figure examines the output gap quantities for the large suppliers more closely by showing the information from Figure 74 separately for each of the three largest QSEs.

Figure 75: Output Gap by Load Level and Participant Size 2005

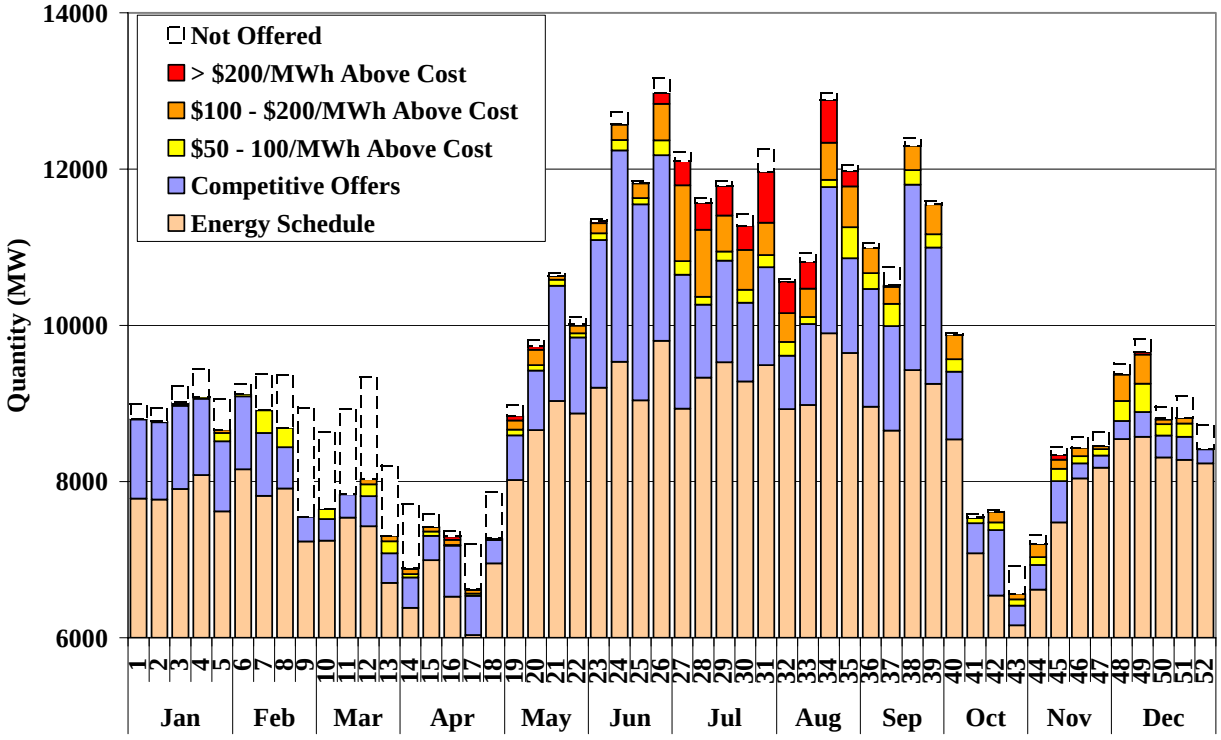


The output gap quantities shown in Figure 75 indicate that each of the three largest suppliers exhibited a similar overall pattern in 2005. Like the small suppliers shown in the previous figure, the three large suppliers' output gap values were correlated with load. This figure shows that the output gap levels were similar in magnitude for Companies A and B, and the small suppliers as a group. However, Company C exhibited larger output gap quantities for both its on-line and off-line units. Company C's average output gap levels exceeded 10 percent of its portfolio, well above 1 GW, when loads exceeded 54 GW. Given the large size of Company C's

portfolio, the pattern observed in the figure above raises significant concerns regarding economic withholding in the balancing energy market. Moreover, to the extent that Company C's output gap is the result of withholding, it is large enough to have substantially impacted market outcomes, particularly under high load conditions. The output gap results in the previous figure warrant a more detailed examination of Company C's conduct during 2005.

In the following analysis, we examined the competitiveness of the balancing energy offers from Company C's portfolio. As a benchmark, we used the Resource Category Generic Fuel Costs ("generic costs") from the ERCOT Protocols to estimate the marginal costs of on-line resources in Company C's portfolio. On-line resources that were not used to provide energy were ranked from lowest to highest generic cost, excluding resources necessary to support their ancillary services obligations. The generic costs of these available resources were compared with the offer prices for each point on Company C's up-balancing offer curve. The results of this comparison are shown below in Figure 76.

Figure 76: Comparison of Offer Prices and Generic Marginal Costs
Weekly Average for the Afternoon Hours in 2005
Offers from Company C



The figure shows that Company C's energy schedules and offer quantities exhibit the typical seasonal pattern of rising in the winter, decreasing in the spring and fall, and peaking in the summer. With the exception of a period from February through April, Company C offered virtually all of its available capacity to the balancing market. For this analysis, we considered offers competitive when the offer price was within $50 per MWh of the generic cost of the underlying resource. During the summer months, Company C typically scheduled 9 to 10 GW of energy and offered 2 to 4 GW into the balancing market, leaving only a small quantity of un-offered balancing energy. However, in the final week of June, Company C began to offer large quantities of energy at prices far in excess of its generic costs. From the last week of June through the end of August, the pattern continued with Company C averaging more than 1 GW of offers priced more than $50 per MWh above generic costs. It should also be noted that the previous analysis indicated Company C had substantial output gap quantities associated with its off-line units as well.

The results of the analysis above raise significant concerns regarding economic withholding. According to the output gap results, Company C had a larger share of unutilized economic capacity than the other suppliers examined in this section, despite the fact that it offered almost all of its available energy in the balancing energy market during most of the year. Moreover, this potentially withheld capacity grew at higher load levels when the ability to raise prices is generally highest. We did not analyze the impact on balancing market prices for this report. Nonetheless, Company C's balancing energy offer patterns raise substantial competitive concerns.

APPENDIX A

Figure 77: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to Houston – 2005

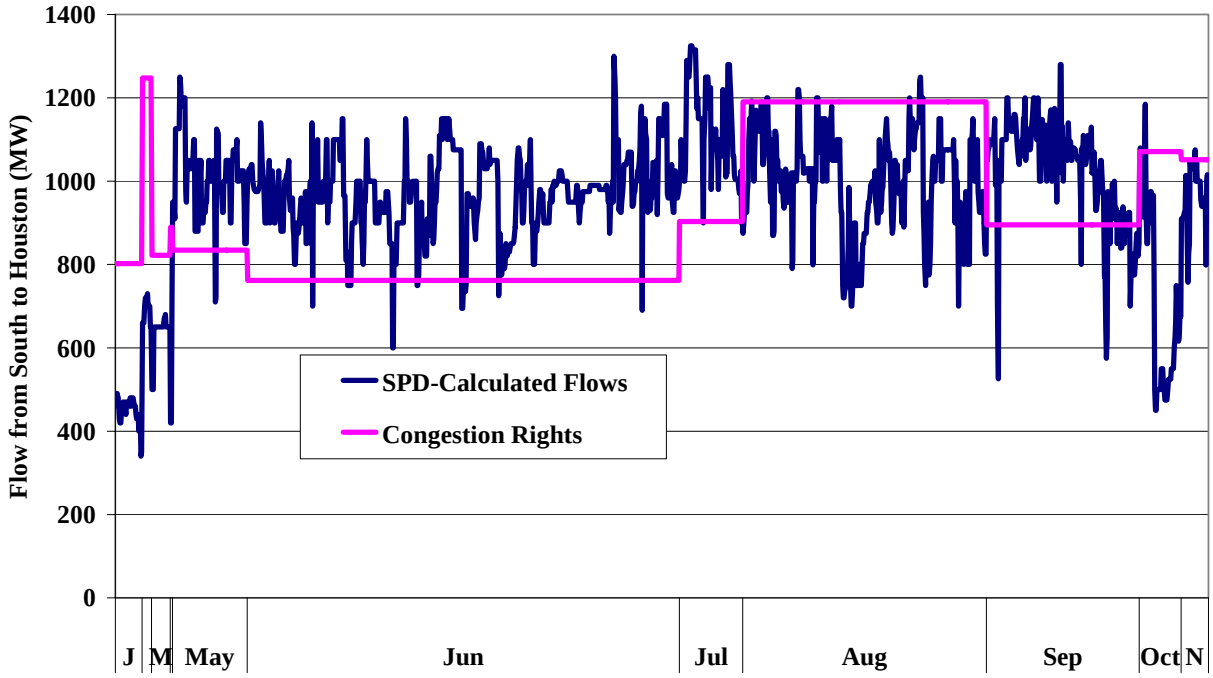


Figure 78: Congestion Rights Allocated vs. SPD Flows During Constrained Intervals North to Houston – 2005

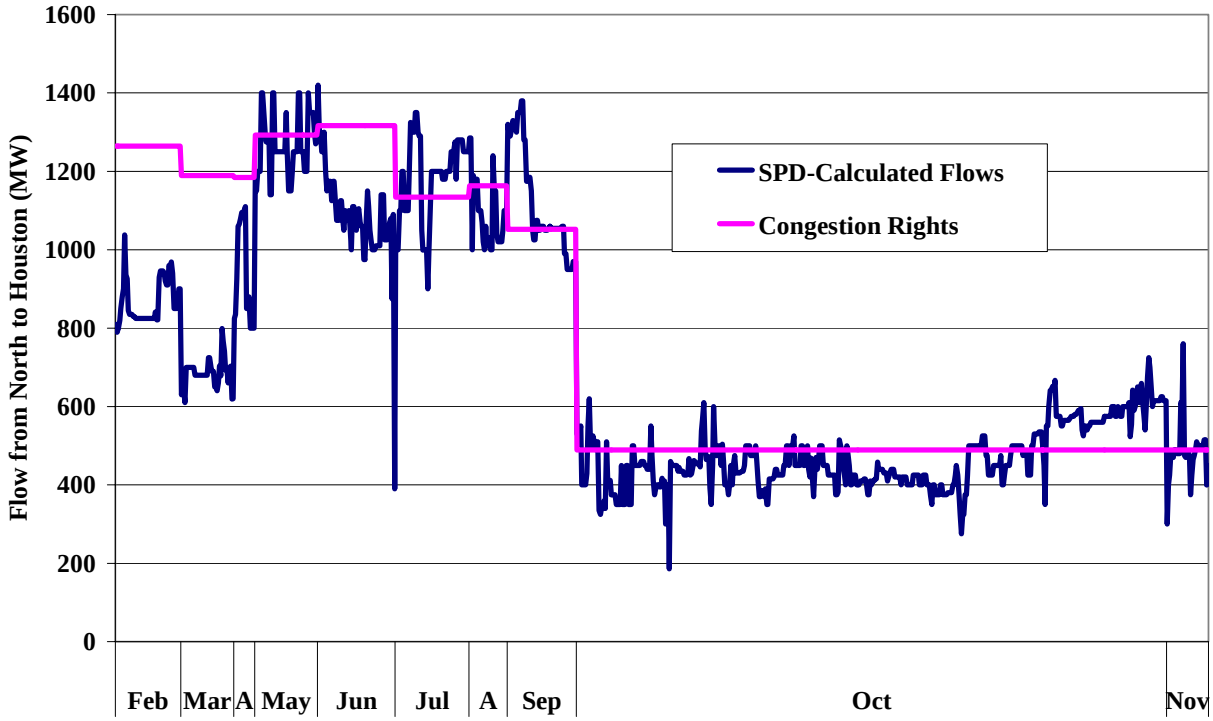


Figure 79: Congestion Rights Allocated vs. SPD Flows During Constrained Intervals North to West – 2005