month, along with the net purchases or sales (*i.e.*, balancing up energy minus balancing down energy).

Figure 12: Average Quantities Cleared in the Balancing Energy Market 2002 to 2006

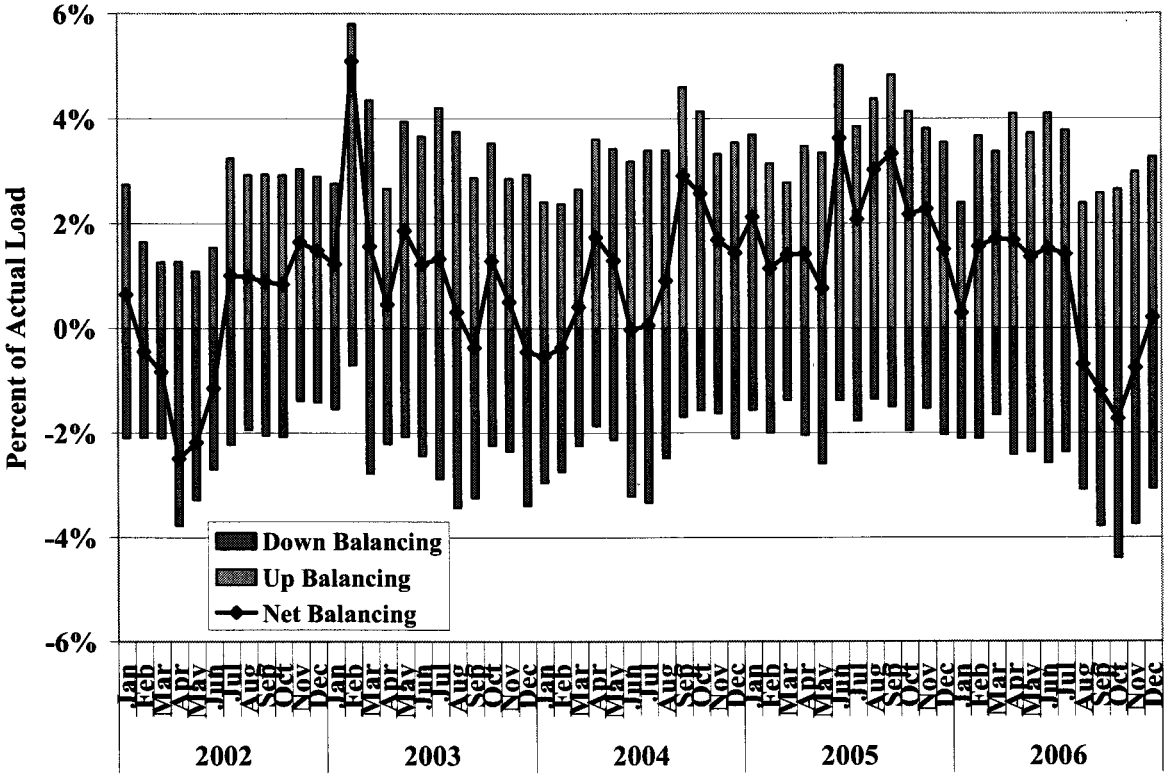


Figure 12 shows that the total volume of balancing up and balancing down energy as a share of actual load increased from an average of 4.6 percent in 2002 to 6.1 percent in 2003, 5.7 percent in 2004, 5.6 percent in 2005, and 6.1 percent in 2006. Thus, there was a general increase in trading through the balancing energy market after 2002. Over time, the volume of balancing up energy has risen relative to the volume of balancing down energy. However, starting in August 2006, the average volume of balancing down energy began to increase. In 2006, the average amount of net balancing up energy (*i.e.*, balancing up minus balancing down) was 1.3 percent. 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 capacity exists to supply the need, but is not available in the 15-minute timeframe 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 13 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 13 shows the portion of the hours during 2006 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 14 percent of the hours in 2006.

Figure 13: Magnitude of Net Balancing Energy and Corresponding Price 2006

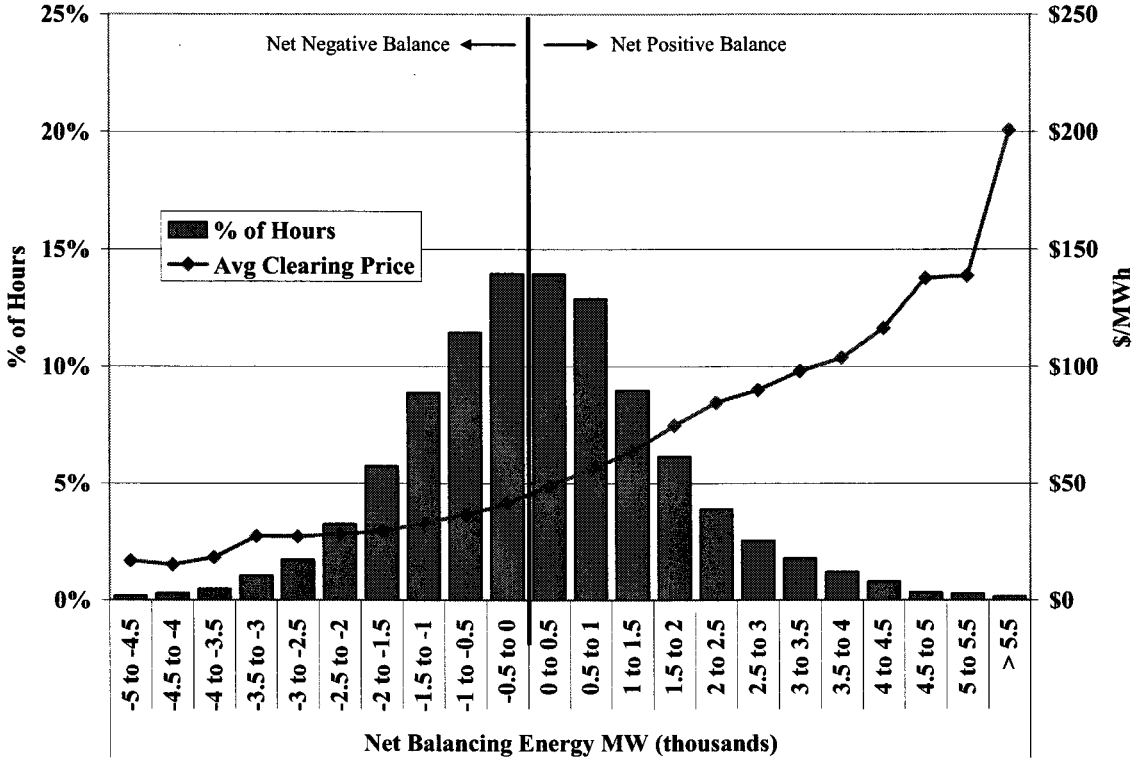


Figure 13 shows a relatively symmetrical distribution of net balancing energy purchases centered around zero gigawatts. This is consistent with Figure 12 which showed that there were comparable portions of net balancing up and down quantities on average during 2006. In approximately 52 percent of the hourly observations shown, Figure 13 also shows that net balancing energy schedules averaged between -1.0 and 1.0 gigawatts.[10] Hence, there were many hours when the net balancing energy traded was relatively low, because the total scheduled energy was frequently close to the actual load.

The line plotted in Figure 13 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,

10 One gigawatt corresponds to roughly 3 percent of the average actual load in ERCOT.

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 2004 and 2005.[11] 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

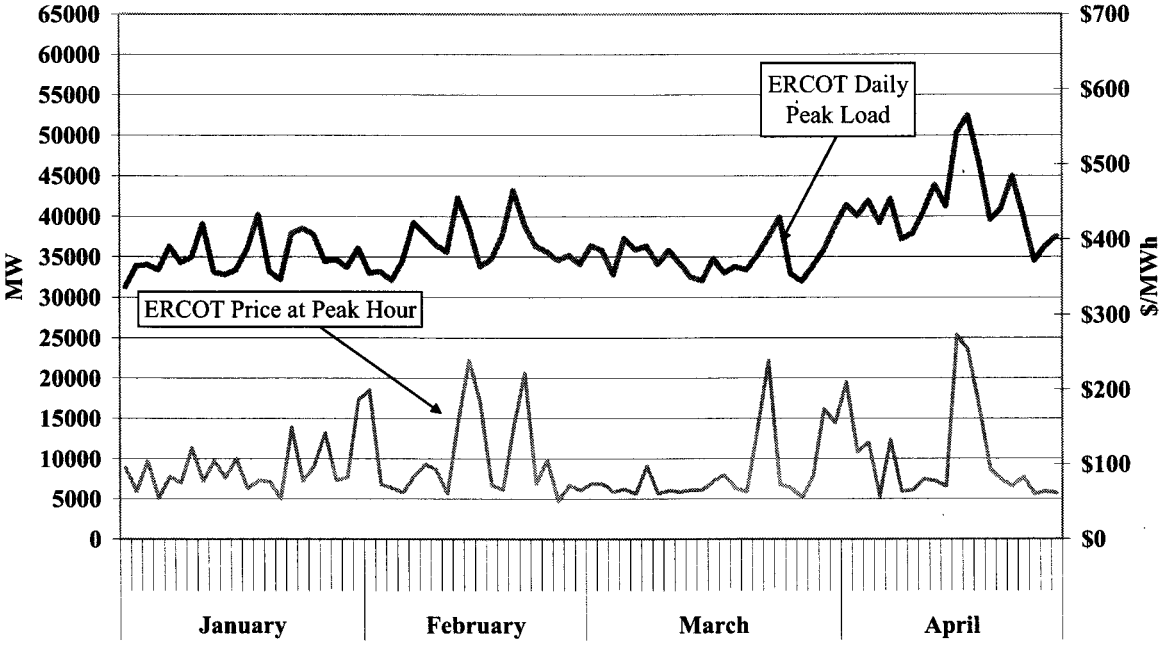
The prior section shows that the level of net sales in the balancing energy market appears to play a significant role in explaining the balancing energy prices. In this section, we examine this relationship in more detail, as well as the role of more fundamental determinants of balancing energy prices,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

Figure 14 shows the average balancing energy price and the actual load in the peak hour of each weekday during 2006.

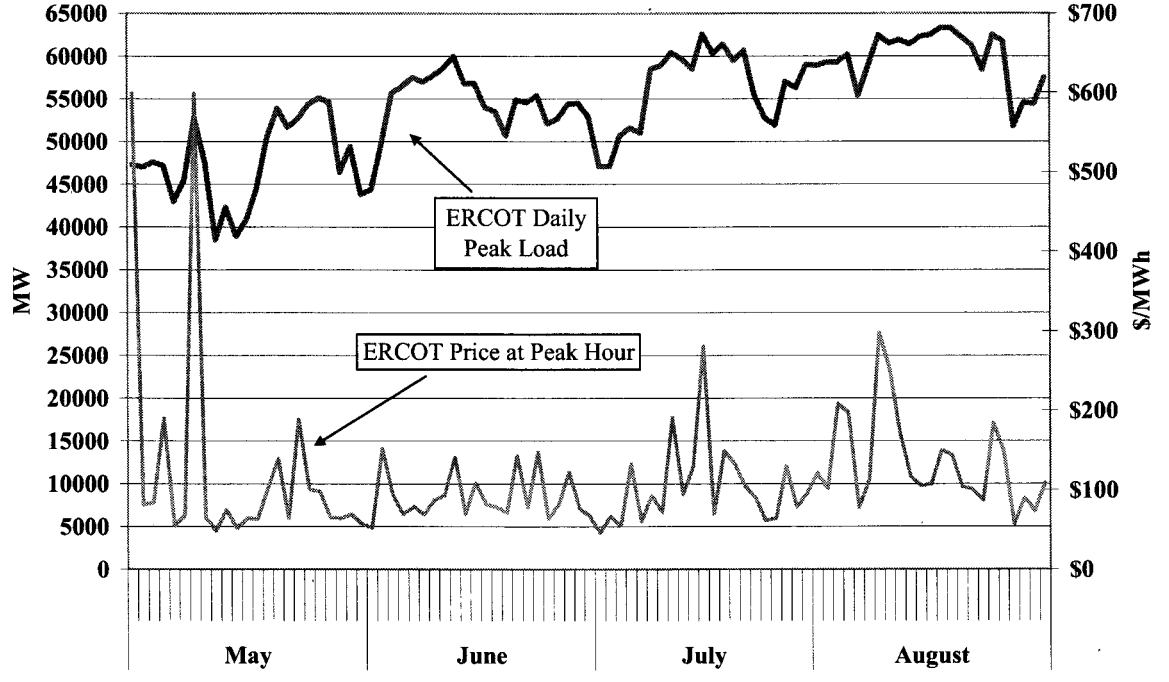
11 See 2004 SOM Report and 2005 SOM Report

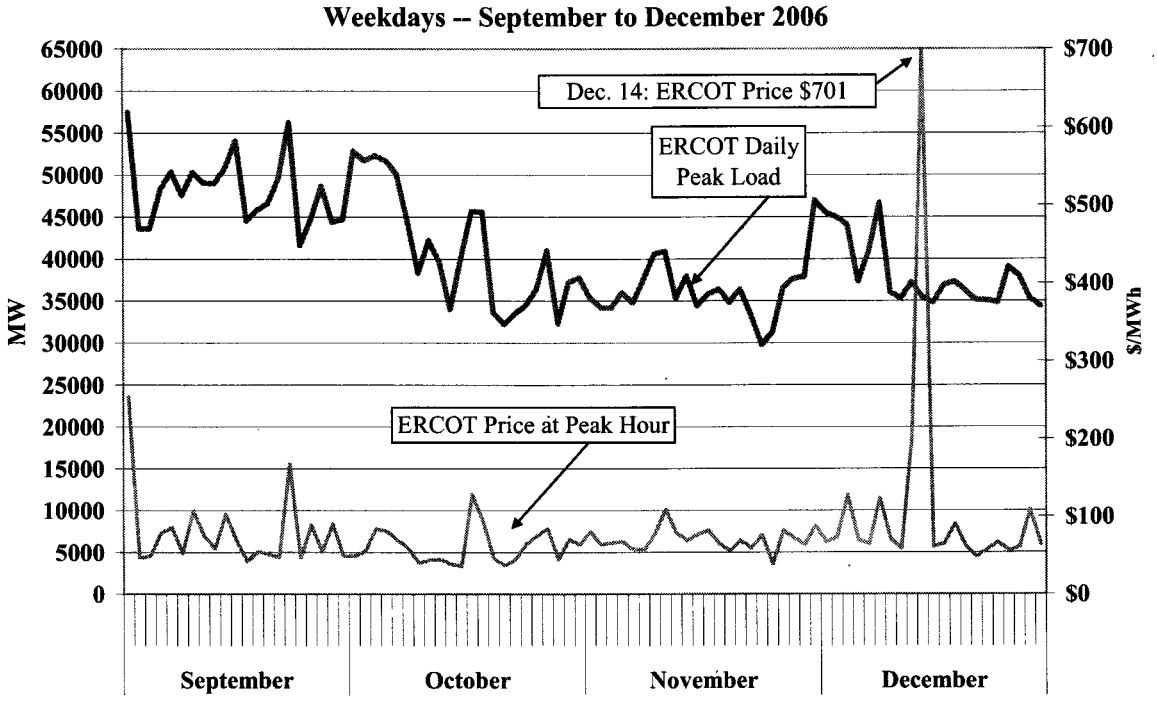
Figure 14: Daily Peak Loads and Balancing Energy Prices

Weekdays -- January to April 2006



Weekdays -- May to August 2006





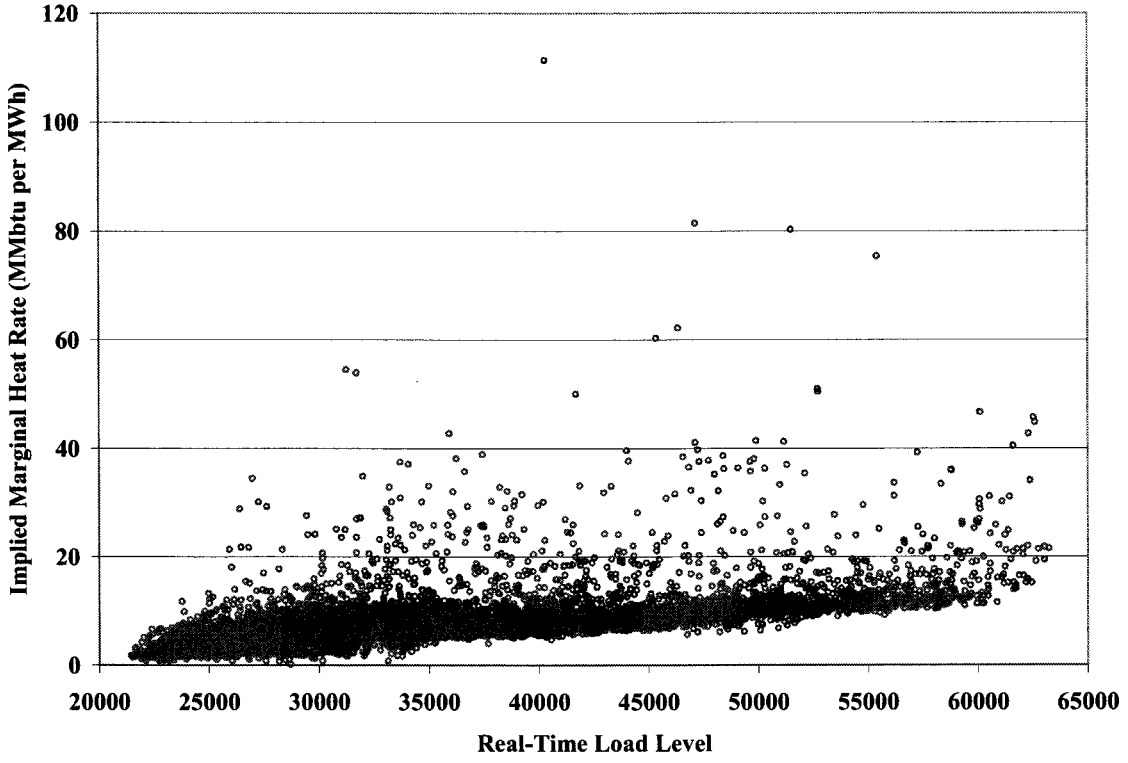
The figure shows that a large share of the days with high prices (*e.g.*, greater than $200/MWh) coincide with periods when demand is high or rising quickly relative to the previous several days. However, prices spikes also occurred during lower demand periods. For instance, on December 14, the price at peak load hour reached $701, while the peak load was lower than the previous 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 an unanticipated rise in load. In ERCOT, prices in the balancing market can reach extremely high levels even when demand is not particularly high. This is primarily due to structural inefficiencies in the balancing energy market that are inherent to the zonal market model, the lack of a centralized unit commitment, load forecast errors, and the fact that the excess online capacity during peak load hours has consistently dropped over the last several years.

To further examine the relationship between actual load in ERCOT and balancing energy prices, Figure 15 shows the hourly average gas price-adjusted balancing energy prices versus the hourly average 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 15: Hourly Gas Price-Adjusted Balancing Energy Price vs. Real-Time Load 2006



The figure indicates a positive correlation between real-time load and the clearing price in the balancing market. Although prices were generally higher at higher load levels, the analysis shown in Figure 13 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 16 shows the average load and balancing energy price in each interval from 4 AM through 1 PM in 2006. The price is plotted as a line in the figure

while the average load is shown with vertical bars.

Figure 16: Average Clearing Price and Load by Time of Day
Ramping-Up Hours – 2006

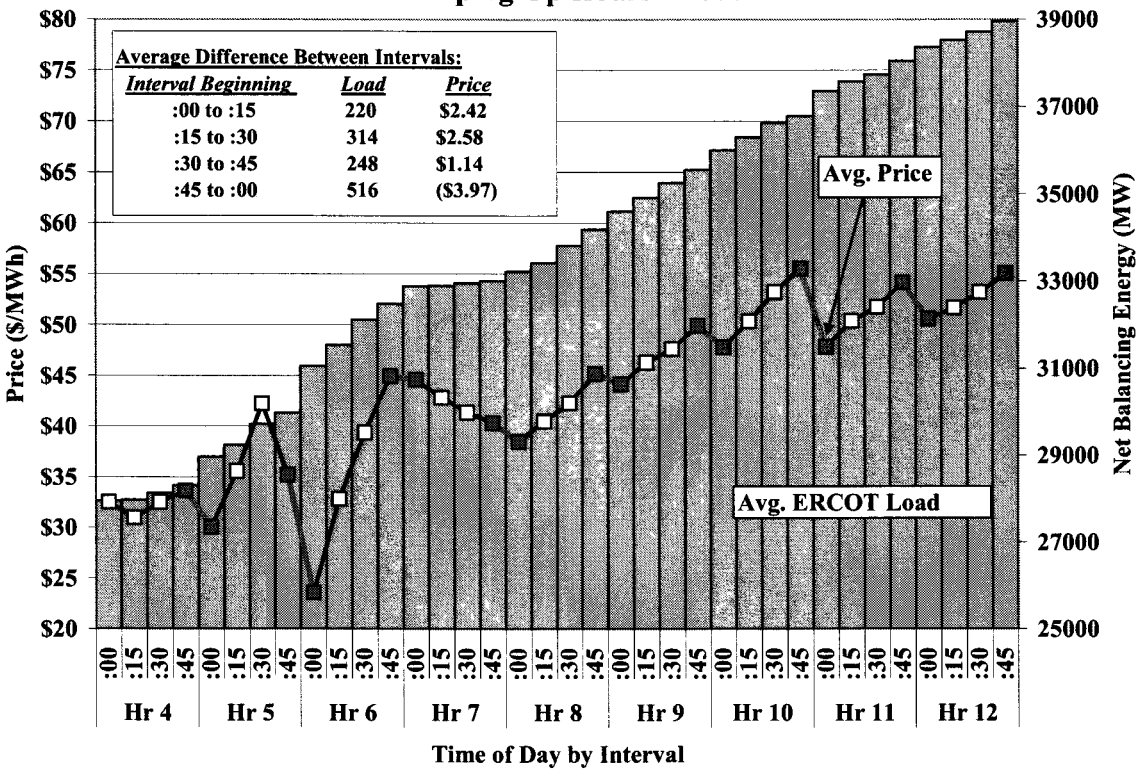


Figure 16 shows that, with the exception of hour 7, the load steadily increases in every interval and prices generally move upward from about $32 per MWh at 4:00 AM to $55 per MWh at 12:45 PM. If actual load were the primary determinant of energy prices, the balancing energy prices would rise gradually as the actual load rises. However, Figure 16 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 -$3.97 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 17 shows this decrease in load by interval, together with the average balancing energy prices for the intervals from 9 PM to 3 AM.

Figure 17: Average Clearing Price and Load by Time of Day
Ramping-Down Hours – 2006

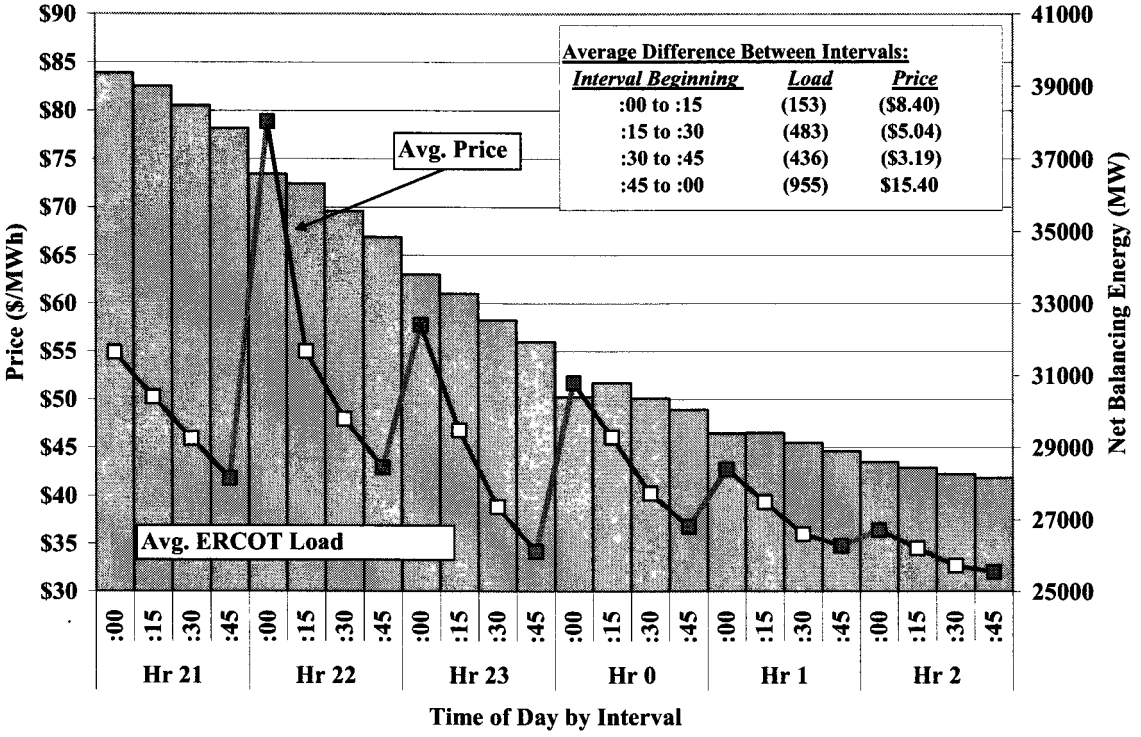


Figure 17 shows that while balancing energy prices decrease over these intervals, they follow a similar pattern as exhibited in the ramping-up hours. The balancing energy price decreases in each interval of the hour before rising substantially in the first interval of the following hour. The balancing energy price increases by an average of $15.4 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 2004 and 2005.[12]

These figures show that this pattern of balancing energy prices by interval is not explained by

12 See 2004 SOM Report and 2005 SOM Report

58

changes in actual load. Rather, changes in balancing energy deployments by interval underlie this pricing pattern. Sizable changes in balancing energy deployments occur between intervals, particularly in the first interval of the hour. These changes are associated with large hourly changes in energy schedules. These scheduling and pricing patterns are examined in detail in Section II below.

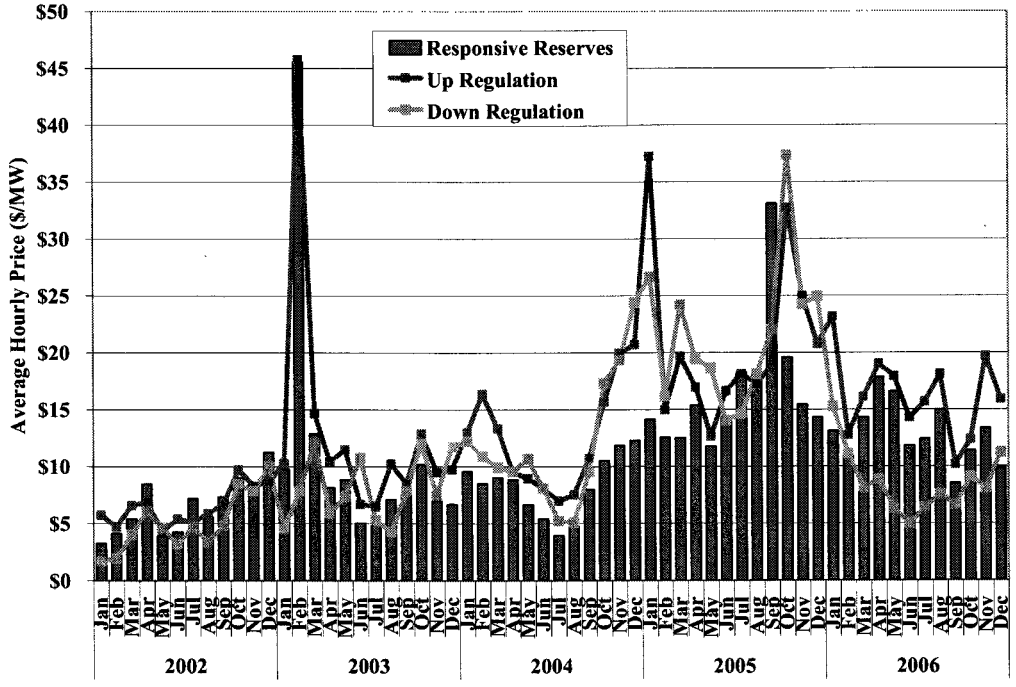
B. Ancillary Services Market Results

The 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 2006.

1. Reserves and Regulation Prices

Our first analysis in this section provides a summary of the ancillary services prices over the past five years. Figure 18 shows the monthly average ancillary services prices between 2002 and 2006. Average prices for each ancillary service are weighted by the quantities required in each hour.

Figure 18: Monthly Average Ancillary Service Prices 2002 to 2006



This figure shows that ancillary services prices have generally risen from 2002 to 2005, but that the price levels moderated in 2006. Much of these price movements can be attributed to the variations 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. In 2005, the pattern shifted such that regulation down prices were four percent higher on average than regulation up prices. However, in 2006, regulation down prices were significantly lower than regulation up prices.

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.

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 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, 23 percent of hours during 2005, and 20 percent of hours during 2006.

Table 1: Average Hourly Responsive Reserves and Non-Spinning Reserves Prices During Hours When Non-Spinning Reserves Were Procured 2002 to 2006

	2002	2003	2004	2005	2006
Non-Spin Reserve Price	$14.51	$9.85	$6.83	$25.10	$21.75
Responsive Reserve Price	$9.20	$10.73	$9.10	$28.16	$25.55

Table 1 shows that responsive reserves prices are higher on average than non-spinning reserves prices during hours when non-spinning reserves were procured. 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 that must be delivered in 10 minutes from on-line resources while non-spinning reserves must be delivered in 30 minutes.

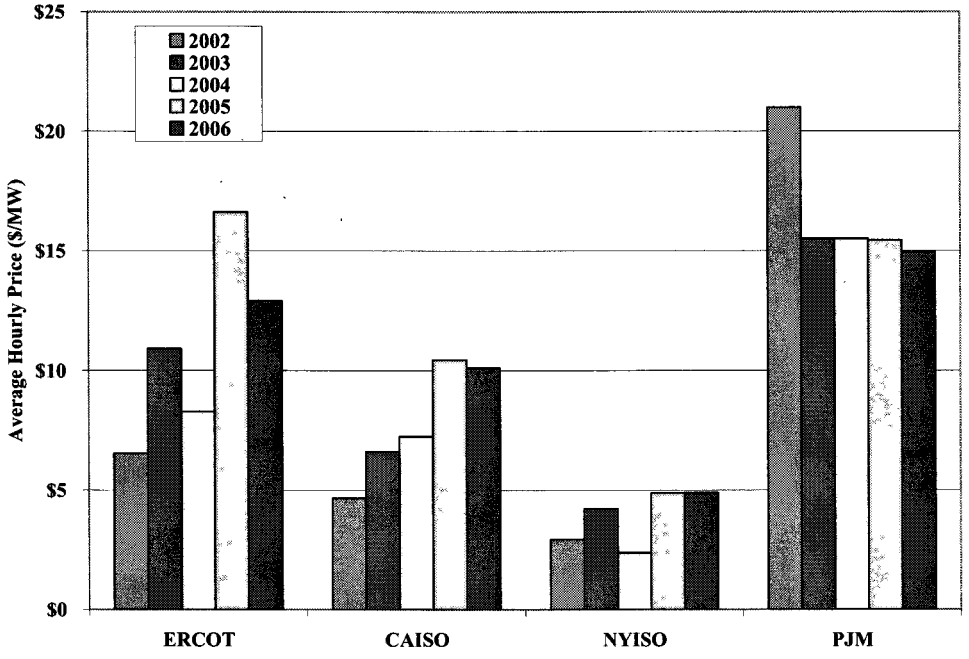
Generators incur two types of costs associated with providing reserves in the ERCOT market. First, reserves providers incur opportunity costs from any profitable sales they forego in the energy market. For generators, this is the same regardless of whether the generator is providing responsive or non-spinning 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 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 often 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 2006, about 2 percent of the responsive reserves were actually deployed, while 5.2 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 responsive and non-spinning reserves is to protect the system against unforeseen contingencies (*e.g.*, generator outages or load forecast error), 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 demand is unusually high or unpredictable or the resources projected to be available in real-time may not be sufficient to satisfy the energy demand while meeting the responsive and regulation up reserve requirements, ERCOT will procure 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 if needed. ERCOT always procures 2,300 MW of responsive reserves to ensure adequate protection against the loss of the two largest units.

Responsive reserve prices dropped in 2006 from 2005, but remained higher than the prices observed in 2002 to 2004. Figure 19 shows how the annual average prices in ERCOT from 2002 to 2006 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, but lower than PJM during 2006.

Figure 19: Responsive Reserves Prices in Other RTO Markets 2002 to 2006



62

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, additional regulation resources are 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 20 shows the relationship between the quantities of regulation required by ERCOT and regulation price levels. This figure compares regulation prices to the average regulation quantity (both up and down regulation) 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 1,280 MW of regulation capability prior to the initial ramping period (beginning at 6 AM). The requirement then jumps up to about 2,000 MW during the steepest ramping hours from 6 AM to 9 AM. The requirement declines to about 1,500 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 1,970 MW.

Figure 20: Regulation Prices and Requirements by Hour of Day 2006

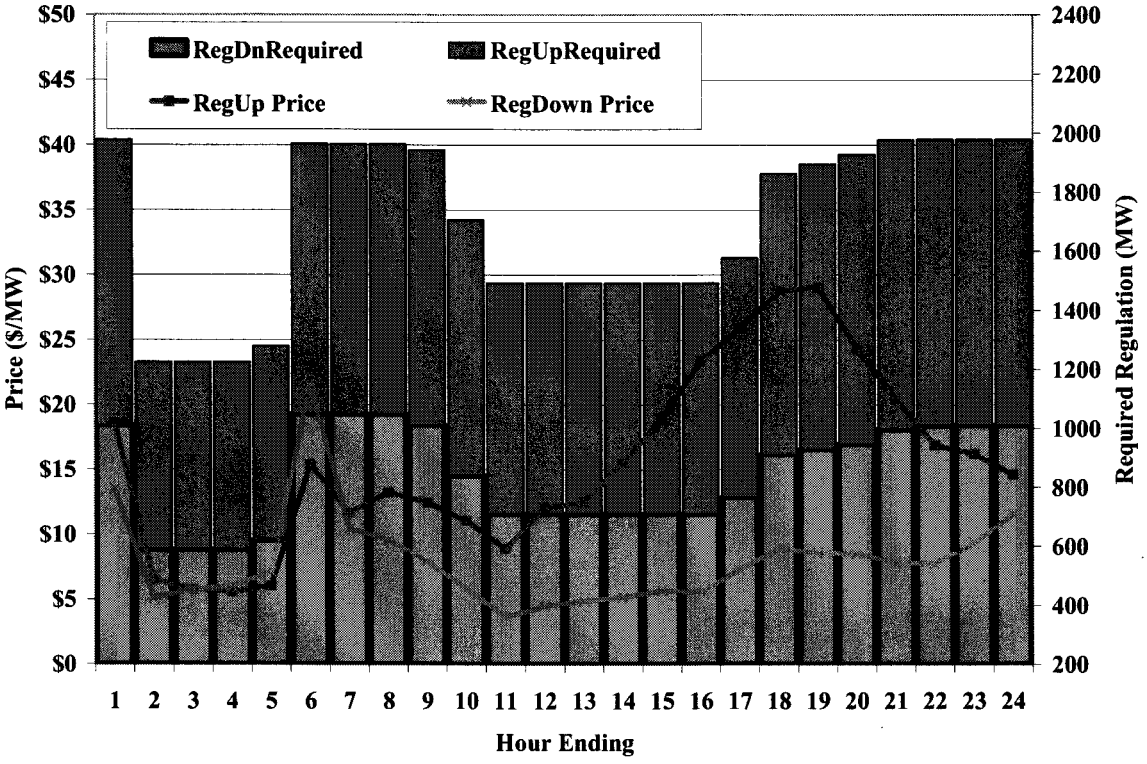


Figure 20 indicates that average regulation prices are generally correlated with the regulation quantity purchased and the typical load pattern in ERCOT. During non-ramping hours, such as overnight and late morning, regulation up and down prices range from $5 to $15 per MW. During the ramping hours in early morning and evening, average regulation up and down prices range from $10 to $20 per MW. In the afternoon hours, regulation up prices range from $20 to $30 and regulation down prices range from $5 to $10 per MW. Regulation up prices are higher on average in the afternoon hours because load levels and balancing energy prices are typically higher in these hours and the amount of capacity available to supply regulation up is lower than in other hours.

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 21 summarizes the average amounts of regulation procured through the auction and/or bilateral arrangements on an annual basis since 2002.

Figure 21: Annual Average Regulation Procurement 2002 to 2006

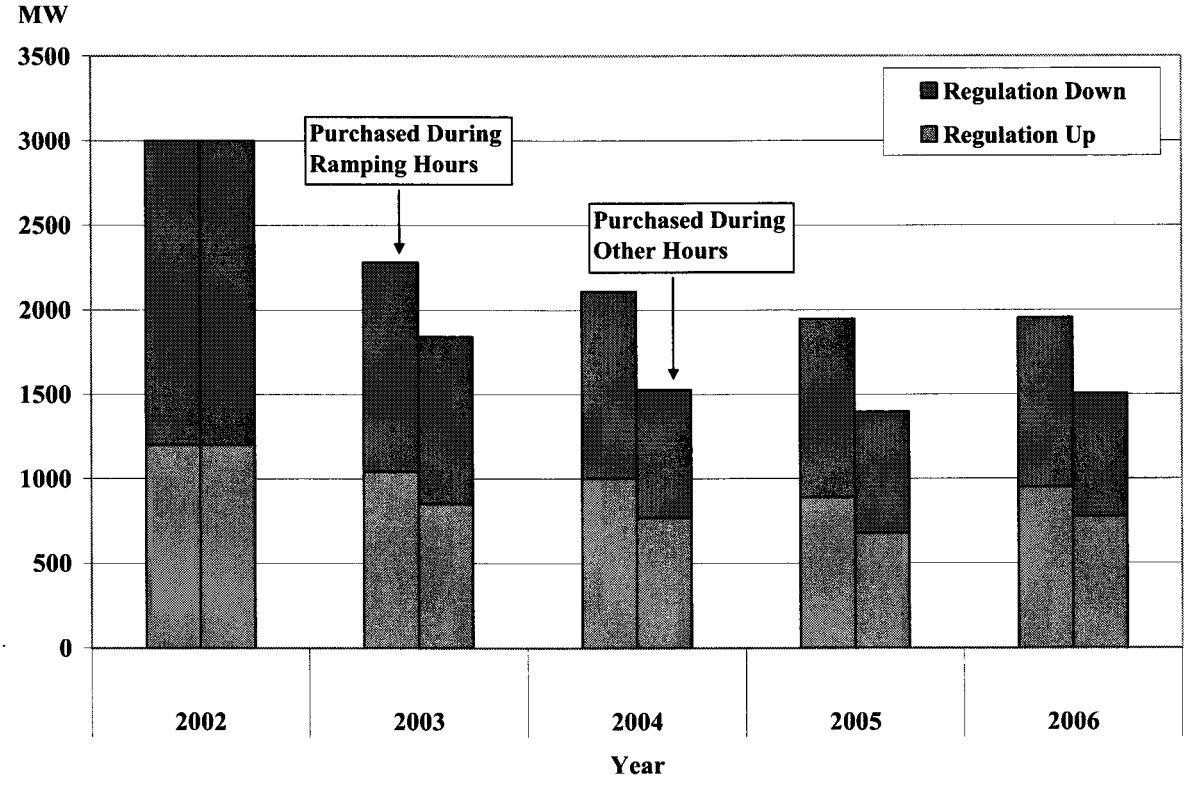


Figure 21 shows that ERCOT has reduced the average regulation quantity scheduled since 2002. The largest reduction was from 2002 to 2003, although the reductions in the remaining two years were also substantial. The regulation quantities required in 2006 was almost the same as in 2005 during ramping hours and was a slightly higher quantity than 2005 during non-ramping hours. Overall, ERCOT has lowered the required amount by 35 percent during ramping hours and 50 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 2006.

Currently, ERCOT's regulation procurement methodologies group regulation procurement quantities into 4 to 6 blocks of hours and procure the same quantity in each block for each day in each month. In late 2006, the Independent Market Monitor ("IMM") initiated discussions with ERCOT to investigate modifications to this methodology that would allow for a different quantity of regulation to be procured in each hour of each day during a month based upon analysis of historical deployment data. The ERCOT Board approved the changed methodology in June 2007 to be implemented in August 2007. It is expected that this change will reduce the overall quantities of regulation procured over all hours, but may increase the regulation quantities procured in certain hours. This change should result in more efficient procurement of regulation up and down service while maintaining or even improving reliability.

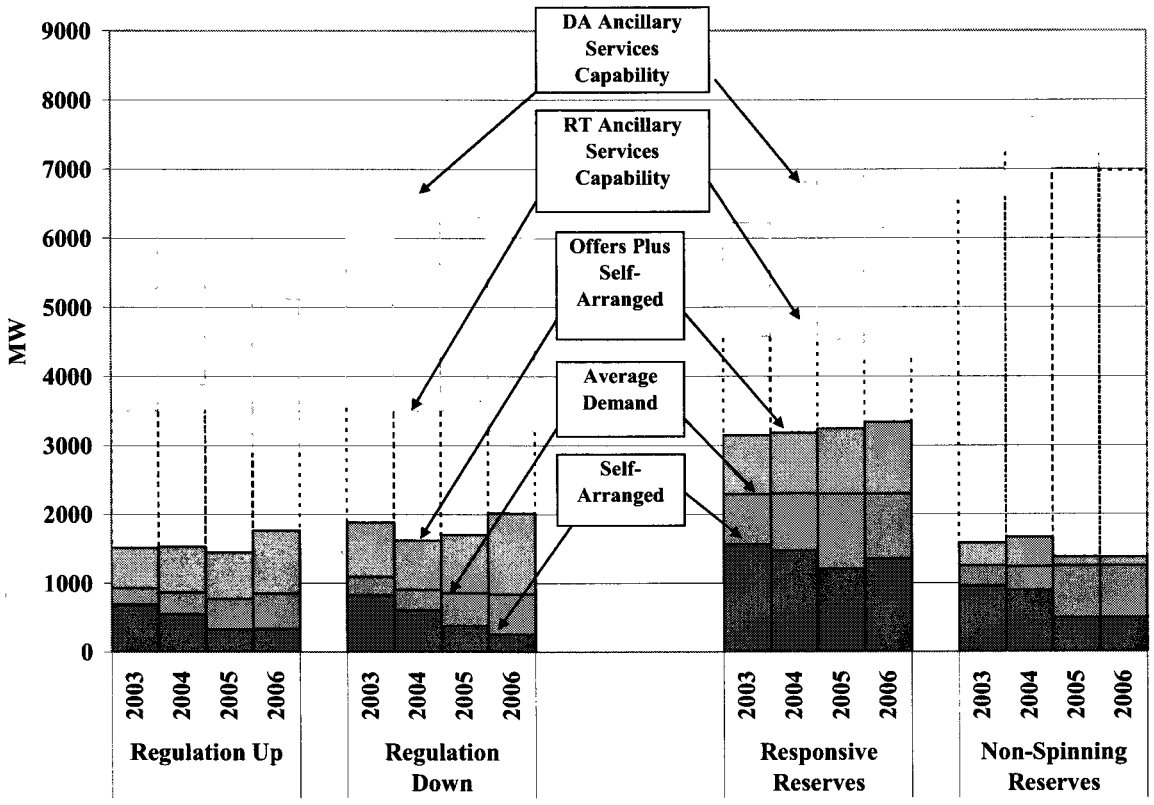
2. Provision of Ancillary Services

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 22. 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 22 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 22: Reserves and Regulation Capacity, Offers, and Schedules 2003 to 2006



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 22 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 22 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 22 includes the capability of units that QSEs indicate are able to ramp-up in thirty minutes and able to start-up on short notice. 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 22 shows that except for responsive reserve in 2006, in which about 54 percent of available responsive reserve capacity was offered, less than one-half of each type of ancillary services capability was offered during 2003, 2004, 2005, and 2006. 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.

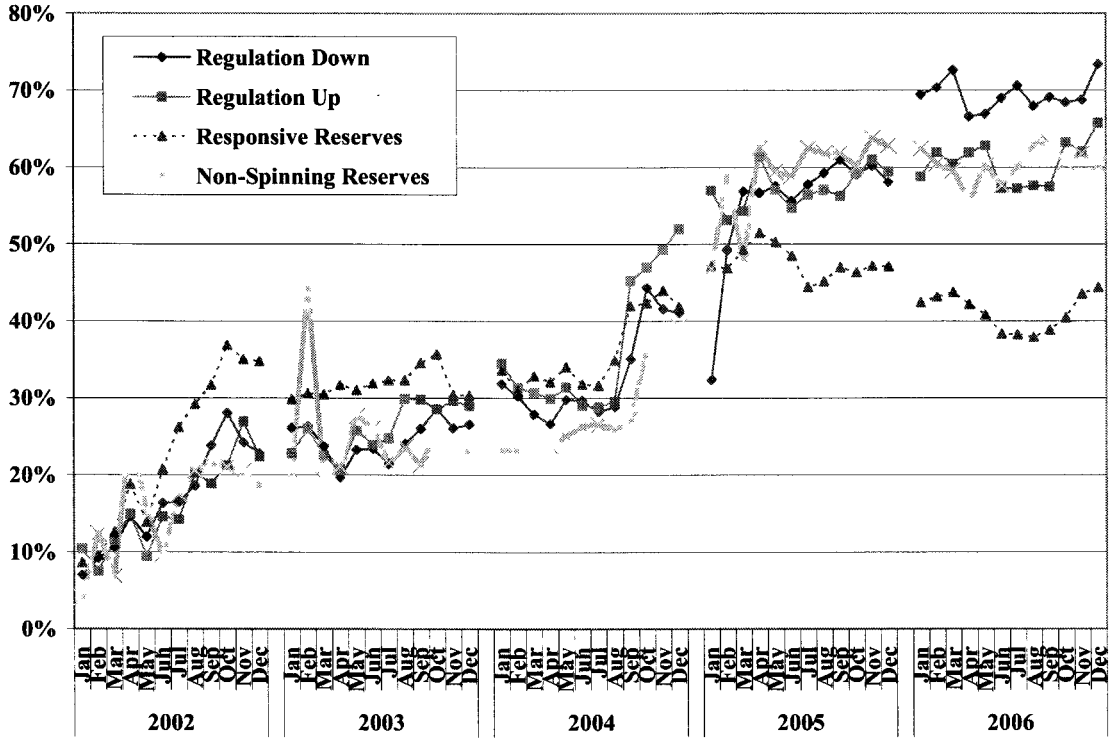
In addition, participants may not offer the capability of resources they do not expect to commit for the following day. Suppliers could submit offer prices high enough to ensure that their costs of committing additional resources to support the ancillary services offers are covered. However, under the current market design, ancillary services are procured independently for each hour and not optimized over the entire day (e.g., including minimum run times and minimum quantities), which greatly increases the risk associated with this approach. The nodal market will include co-optimized procurement of energy and reserves over the entire operating day, which should enhance the efficiency of the procurement of reserves. On average, there is often a substantial quantity of reserves that remain available in real time, but that is 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 22 shows modest changes in the amount of day-ahead ancillary services capability between 2003 and 2006. 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 each year 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 23 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 23 shows the share of each type of ancillary service that is purchased through the ERCOT market.

Figure 23: Portion of Reserves and Regulation Procured Through ERCOT 2002 to 2006



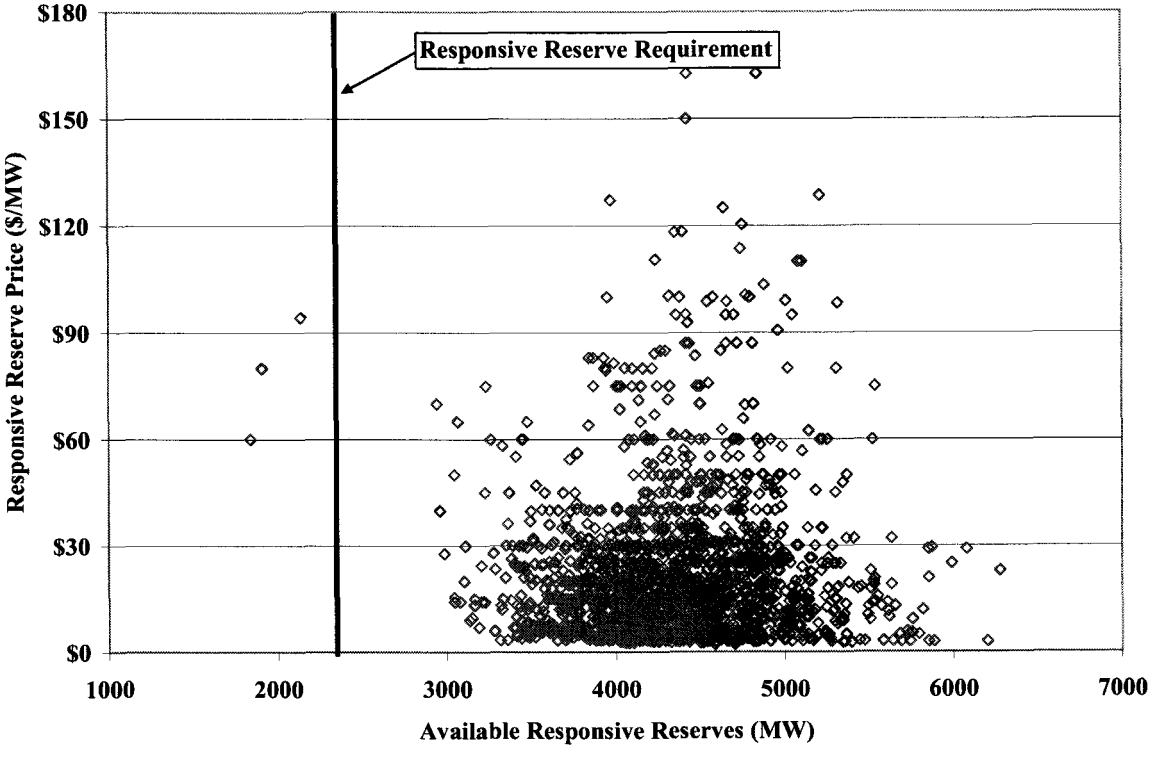
This figure shows that purchases of all ancillary services from the ERCOT markets have generally increased over time, although the purchases of responsive reserve from the ERCOT

market has dropped slightly over the last two years (*i.e.*, the quantity of self-arranged responsive reserve has increased slightly over the last two years). 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 2006. Prices in this market are significantly higher than in other markets that co-optimize the procurement and dispatch of energy and responsive reserves. Lower prices occur in co-optimized markets because in the procurement is optimized with energy over the entire operating day and 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 24 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 and, consistent with ERCOT rules, the responsive reserve that can be provided by each generating unit is limited to 20 percent of the unit's maximum capability. The figure also shows the responsive reserves requirement of 2,300 MW to show the amount of the surplus in each hour.

Figure 24: Hourly Responsive Reserves Capability vs. Market Clearing Price Afternoon Peak Hours – 2006



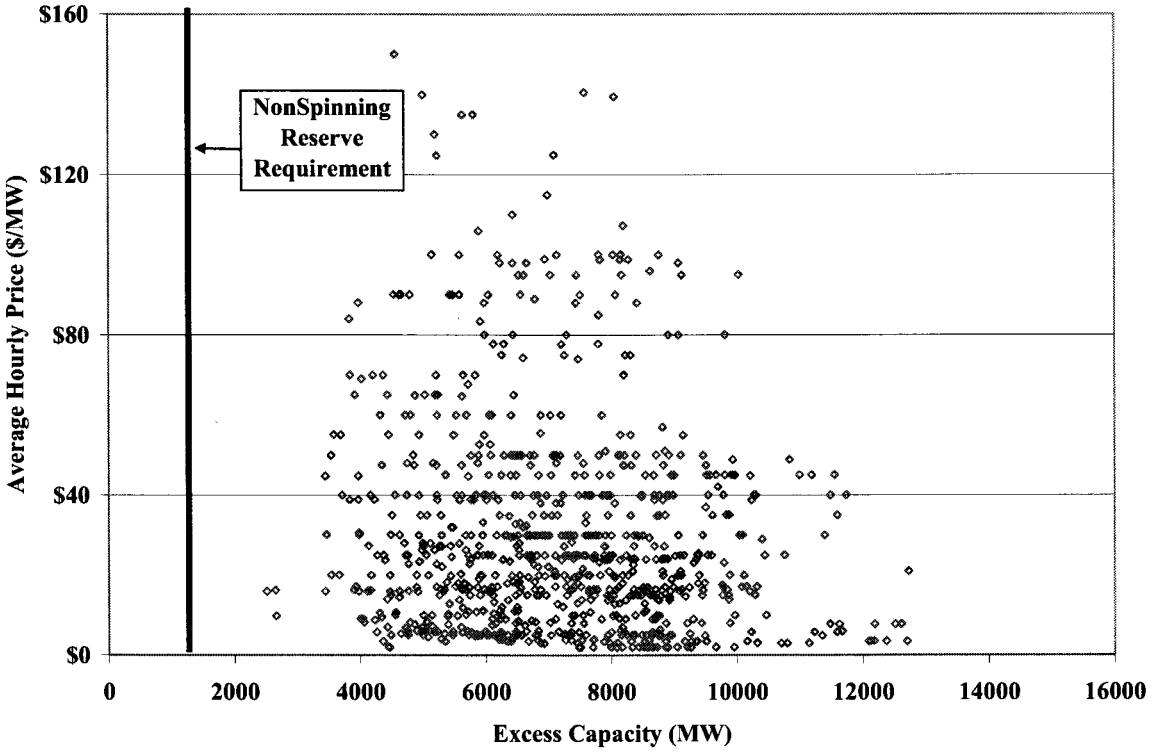
This figure indicates very little relationship between the hourly available responsive reserves capability and the responsive reserves prices in real time. In a well functioning-market for responsive reserves, we would expect excess capacity to be negatively correlated with the clearing prices, but this was not the case in 2006. Similar analyses in previous reports show the same lack of correlation between prices and available reserves. These results reinforce the potential benefits promised by jointly optimizing the operating reserves and energy markets, which is currently being developed for implementation in the nodal market by 2009 (day ahead co-optimization, but not real-time).

Non-spinning reserves are purchased on a day-ahead basis primarily during defined times of extreme or unpredictable demand. Non-spinning reserves are resources that can be deployed 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 25 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 2006.

Like the previous analysis of responsive reserves, the results shown in Figure 25 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 2004 and 2005. In a well functioning-market for non-spinning reserves, we would expect excess capacity to be negatively correlated with the clearing prices.

Figure 25: Hourly Non-Spinning Reserves Capability vs. Market Clearing Price All Hours 2006

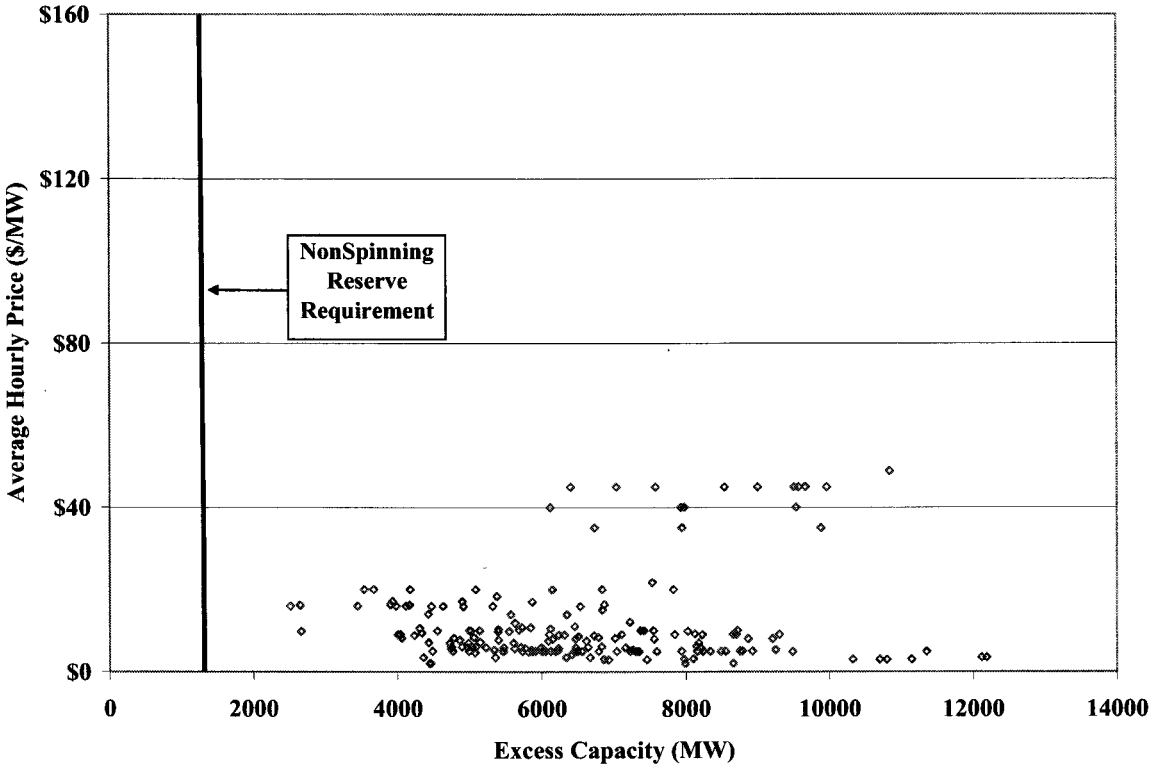


One factor affecting non-spinning reserve prices is that, prior to September 2006, the deployment of non-spinning reserves typically resulted in a significant reduction in the market clearing price of energy. Hence, units deployed for non-spinning reserves would often receive a price for the deployed energy that was significantly less than the operating cost of the unit. In September 2006, new pricing rules were implemented that provide for the recalculation of the energy clearing price when non-spinning reserves are deployed on an *ex post* basis by re-running the

market clearing engine under the assumption that the energy from the deployed non-spinning reserves was unavailable.[13]

Figure 26 shows the data as in Figure 25 for just the months of September though December, 2006. These results clearly show an overall reduction in the clearing price for non-spinning reserves after the implementation of the new rules, which was expected given that the new rules significantly reduce the financial operating risk to providers of non-spinning reserve.

Figure 26: Hourly Non-Spinning Reserves Capability vs. Market Clearing Price September - December 2006



Although the implementation of the new pricing rules associated with the deployment of non-spinning reserves have produced the expected results, ideally the pricing adjustments should be performed in real-time instead of after-the-fact to send accurate and timely price signals to both resources and loads. Further, the current re-pricing mechanism is rather extreme in that it effectively assumes that the energy from non-spinning reserve units is offered at the system-wide offer cap. It would be more reasonable to employ an *ex ante* proxy price that is a function of the

[13] These new rules were approved in Protocol Revision Request No. 650.

incremental costs of deploying an off-line gas turbine. However, because of limitations of the current systems, neither of these improvements is feasible under the current market design.

C. Replacement Reserve Service Market

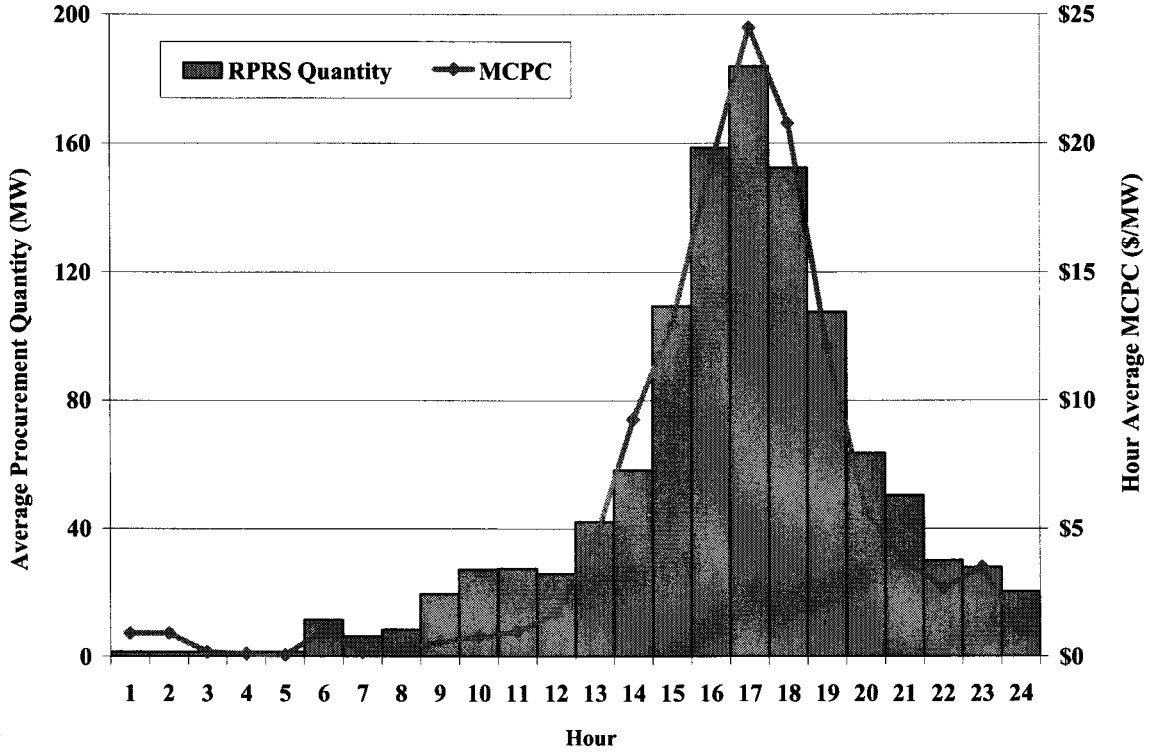
Unit commitment is the day-ahead process of determining which units will be online to meet the forecast demand for the next operating day while observing reliability requirements such as reserve requirements and transmission system limitations. In ERCOT, market participants self-commit units on a decentralized basis to accommodate bilateral energy schedules, ancillary services agreements and load requirements. Building off of this self-commitment, ERCOT conducts a centralized reliability unit commitment to secure additional units that may be necessary to ensure that the system capacity requirement is met and transmission congestion can be resolved in real time operations. Prior to late March 2006, ERCOT relied exclusively upon out-of-merit capacity (OOMC) for this purpose. However, beginning in April 2006, the Replacement Reserve Service (RPRS) market was implemented by ERCOT as the primary tool used to commit capacity in the day ahead to ensure system reliability. Unlike OOMC, RPRS allows ERCOT to optimize unit commitment considering economic and operational factors over all 24 hours of the next operating day.

The RPRS market uses a three step process to commit units and derive the market clearing prices for zonal replacement reserve services. In the first step, the units are selected to satisfy the system load requirement considering transmission limitations (*i.e.*, congestion). Pricing for units selected in step one is cost-based. In the second step, units will be committed when additional capacity is needed to satisfy the forecasted load and system ancillary services requirement. Unlike step one, pricing for units selected in step two is market-based and is a function of the replacement bids submitted by the market participants. Upon the completion of steps one and two, the RPRS market clearing engine generates the market clearing price for each hour for any unit selected in step two. The discussion in this section is limited to replacement reserve quantities procured in step two.

Figure 27 shows the hourly average replacement reserve prices in 2006. As shown in this chart, hour ending 1700 has the highest average market clearing price, which coincides with the typical occurrence of the daily peak load in the summer in the ERCOT market. The market clearing

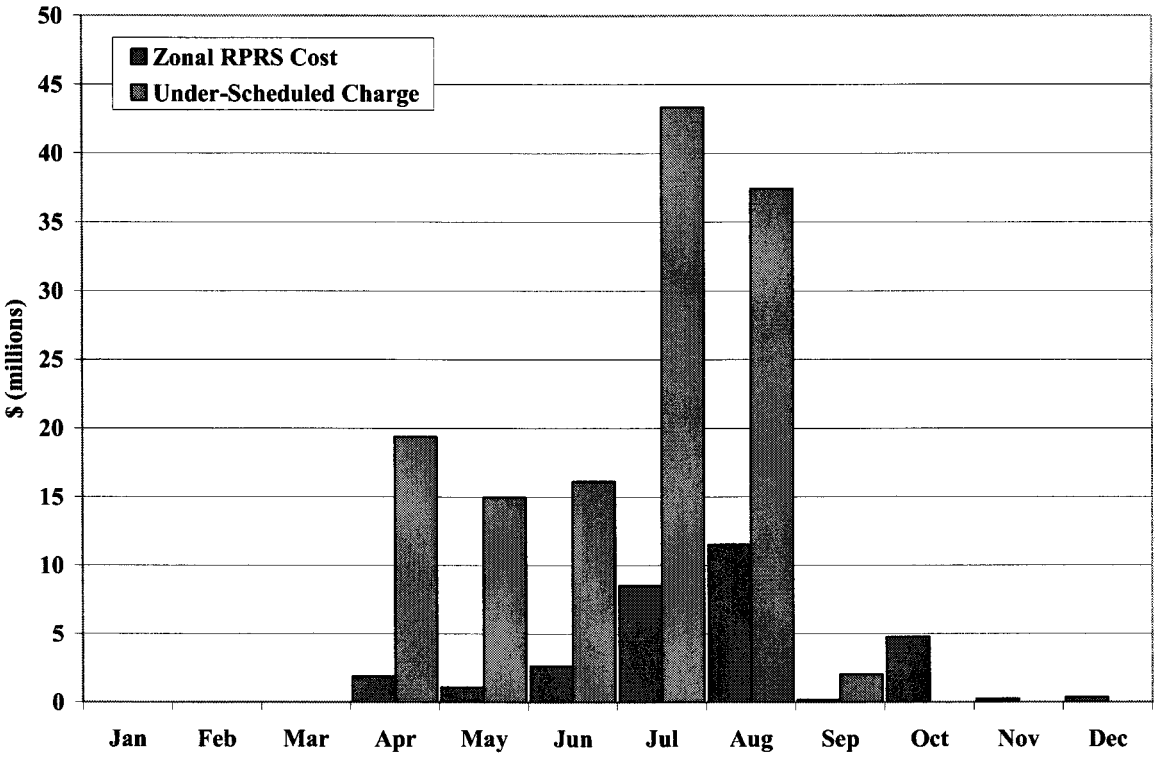
prices for the off peak hours are relatively low, which is consistent with the fact that ERCOT market usually has excess online capacity during off-peak hours.

Figure 27: Replacement Reserve Hourly Average MCPC & Capacity Procurement 2006



From late March through September 2006, costs associated with step two RPRS procurements by ERCOT were directly assigned to QSEs (under-scheduled charge) based upon the RPRS step two clearing price and the measured difference between the day ahead scheduled and actual load of the QSE. Figure 28 shows the zonal RPRS cost and the under-scheduled charge by month for 2006. The under-scheduled charge is greater than the RPRS cost in April through September because the under-scheduled quantity was greater than the quantity of RPRS procured.

Figure 28: Zonal RPRS Cost and Under-Scheduled Charge 2006



Due to concerns raised regarding the accuracy of the cost causation elements associated with the direct assignment provisions of RPRS, the direct assignment provisions were suspended by ERCOT effective October 1, 2006 pending consideration by the PUCT of an appeal relating to these matters.

Ultimately, the PUCT made permanent the suspension of the direct assignment of RPRS step two costs such that all RPRS costs are assigned to all QSEs on a load ratio share for the duration of the existence of the zonal market, noting that the implementation of the nodal market with a centralized day-ahead market and associated provisions related to unit commitment payment and cost allocation should largely resolve the issues associated with the RPRS market in 2006.

D. 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

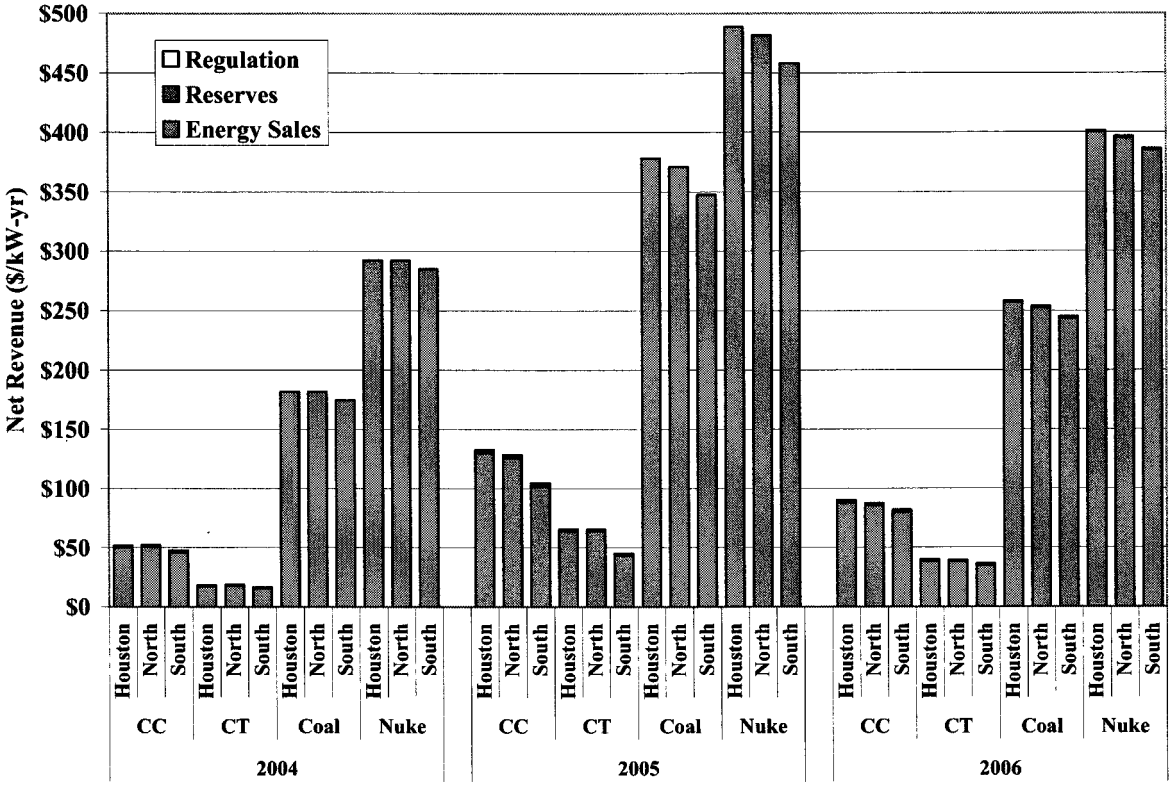
reserves, and regulation markets together provide the economic signals that govern suppliers' decisions to invest in new generation or retire existing generation. In a long-run equilibrium, the markets should provide sufficient net revenue to allow an investor to break-even on an investment in a new 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 low due to mild weather or economic conditions; or
- Market rules are causing revenues to be reduced inefficiently.

Likewise, the opposite would be true if the markets 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 2006 by various types of generators in each zone.

Figure 29 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 29: Estimated Net Revenue 2004 to 2006



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 29 shows that the estimated net revenue for all technologies grew significantly from 2002 to 2003 and again from 2004 to 2005. The net revenue fell in 2006 in each zone compared to 2005; however, net revenue remained higher in 2006 than in years prior to 2005. 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 $85 per kW-year. The estimated net revenue for a new gas turbine in 2006 is approximately $40 per kW-year, which is lower than the estimated net revenue required for new entry. For a new combined cycle unit, the estimated net revenue requirement is approximately $95 to $125 per kW-year. The estimated net revenue in 2006 for a new combined cycle unit is approximately $88 per kW-year, which is also lower than the estimated net revenue required for new entry. 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, return mothballed units to service 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 to support new entry for these technologies. 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 $190 to $245 per kW-year for a new coal unit and $280 to $390 per kW-year for a new nuclear unit. Net revenues were at the lower ends of these ranges in 2003 and 2004, but exceeded them in 2005 and 2006. Thus, it is not surprising that some market participants are expressing interest in

building new baseload facilities in ERCOT.[14] 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.

Although estimated net revenue grew considerably in 2005 and 2006 compared to prior years, 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 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 by 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 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 30 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,[15] 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.[16]

---

14 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.

15 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.

16 The California ISO does not report capacity and ancillary services net revenue separately, so it is shown as a combined block in Figure 30. 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. The California ISO revised its methodology in 2006 to consider a theoretical new combined-cycle unit to participate in both the

Figure 30: Comparison of Net Revenue of Gas-Fired Generation between Markets 2004 to 2006

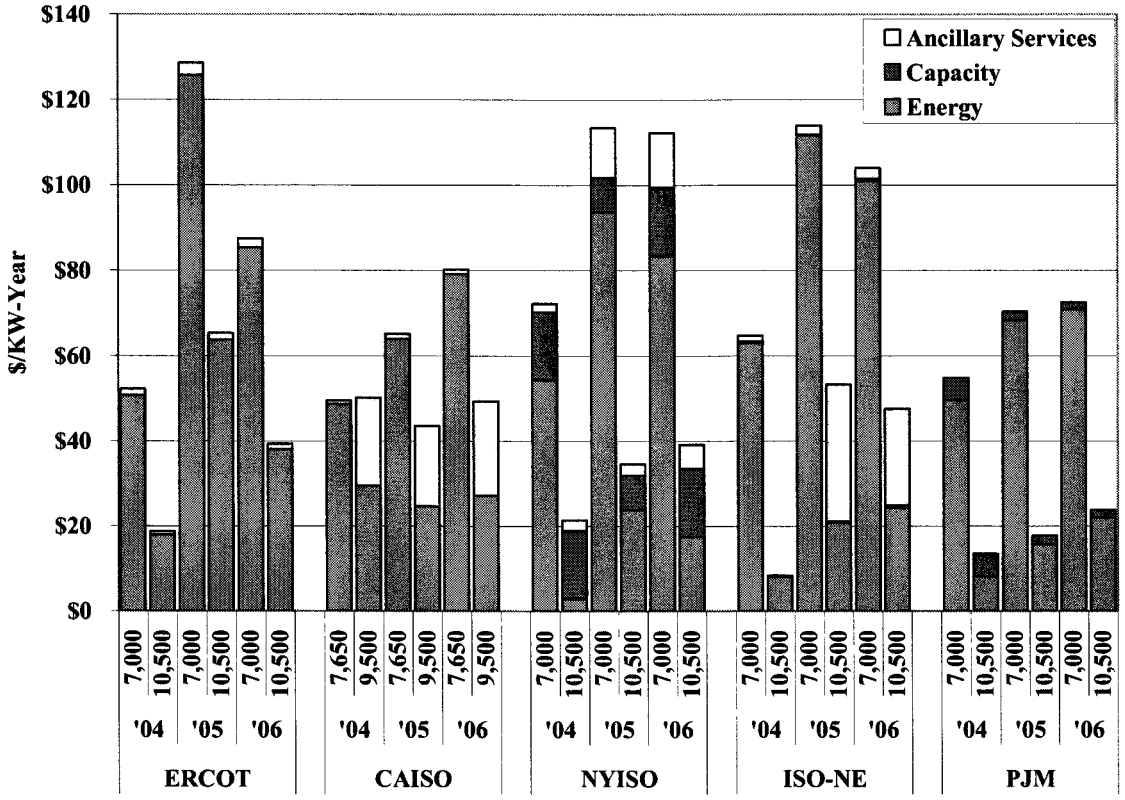


Figure 30 shows that net revenues increased slightly in California, New York and PJM from 2005 to 2006, and decreased in ERCOT and New England. These differences can be explained by several factors. First, ERCOT is much more dependent on natural gas than the other markets. The decrease in natural gas prices in the other regions does not translate as directly into lower 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. In 2006, the New York and New England markets exhibited net revenue in a range that might be sufficient to motivate investment in new gas-fired capacity, while net revenue in ERCOT, 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

Real-time and Day-ahead market, with the net revenues updated from 2004 to 2006.

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. Finally, resource investments are driven primarily by forward price expectations, so historical net revenue analyses do not provide a complete picture of the future pricing expectations that will spur new investment.

The net revenue outcomes in the ERCOT markets in 2006 were primarily affected by the following factors:

- Although continuing to decline relative to prior years, planning reserve margins in 2006 were approximately 16.5 percent, which is well above the minimum requirement of 12.5 percent. Excess capacity lowers net revenue by reducing prices whereas relatively low reserve margins can cause net revenue levels to substantially exceed the annualized cost of a new unit.
- Natural gas prices moderated in 2006, but remained at levels significantly higher than the years prior to 2005. Thus, net revenue for coal and nuclear units continued to be at levels sufficient to support new entry.
- The Modified Competitive Solution Method ("MCSM") triggered price adjustments more frequently in 2006. MCSM is a PUCT-approved mechanism that was in effect in 2005 and through September 2006 that provided for an *ex post* reduction to the resulting market prices when all dispatchable balancing energy was exhausted. The average number of MCSM intervals per month almost doubled to over 26 per month in 2006 compared to less than 16 per month in 2005 for the months in which MCSM was in effect.
- The competitive performance of the ERCOT market improved in 2006.

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:

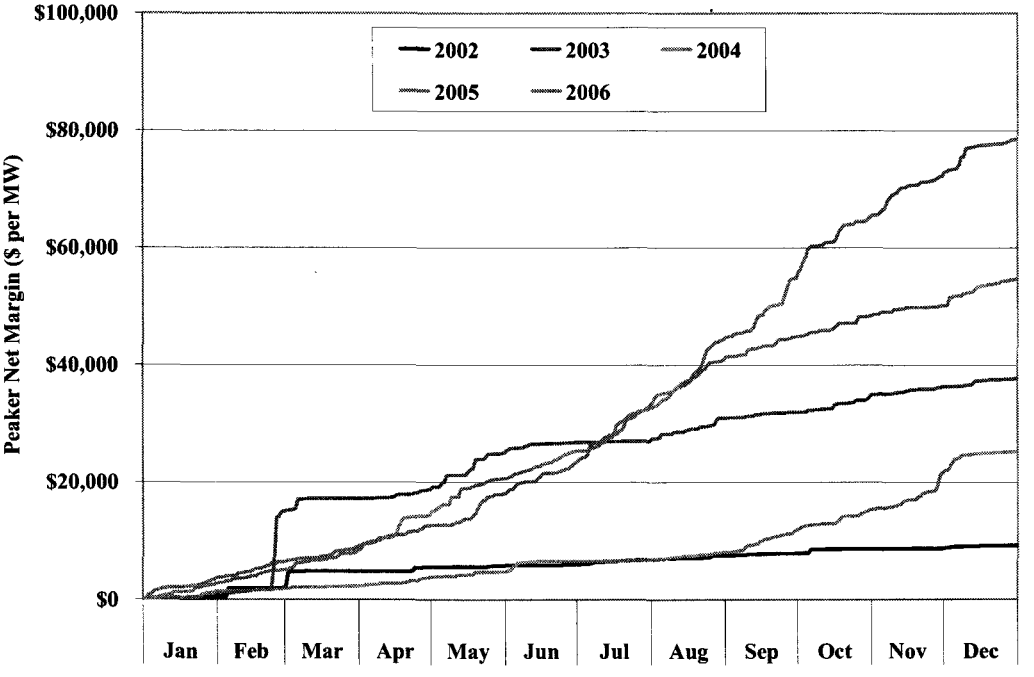
- A capacity market; and/or
- Shortage pricing provisions to ensure that prices rise appropriately in the energy and ancillary services markets to reflect the true costs of shortages when resources are insufficient to satisfy both the energy and ancillary services requirements.

The PUCT adopted rules in 2006 that define the parameters of an energy-only market. These rules include a Scarcity Pricing Mechanism ("SPM") that provides for a gradual increase in the

system-wide offer cap to $1,500 per MWh on March 1, 2007, $2,250 per MWh on March 1, 2008, and to $3,000 per MWh shortly after the implementation of the nodal market. Additionally, market participants controlling less than five percent of the capacity in ERCOT by definition do not possess market power under the PUCT rules. Hence, these participants can submit very high-priced offers. The new rules also eliminated MCSM effective October 1, 2006.

The SPM also includes a provision termed the Peaker Net Margin ("PNM") that is designed to measure the annual net revenue of a hypothetical peaking unit. Under the rule, if the PNM for a year reaches a cumulative total of $175,000 per MW, the system-wide offer cap is then reduced to the higher of $500 per MWh or 50 times the daily gas price index. Although the PNM was not in effect prior to 2007, Figure 31 shows the cumulative PNM that would have been produced for each year from 2002 through 2006.[17]

Figure 31: Peaker Net Margin
2002 to 2006



As previously noted, the net revenue required to satisfy the annual fixed costs (including capital carrying costs) of a new gas turbine unit is approximately $60 to $85 per kW-year (i.e., $60,000

17 The proxy combustion turbine in the Peaker Net Margin calculation uses a heat rate of 10 MMbtu per MWh and includes no other variable operating costs.

to $85,000 per MW-year). Thus, as shown in Figure 31 and consistent with the previous findings in this section relating to net revenue, the PNM reached the level sufficient for new entry in only one of the last five years (2005).

Unlike markets with a long-term capacity market, the objective of the energy-only market design is to allow prices to rise significantly higher during legitimate shortage conditions (*i.e.*, when the supply of resources is insufficient to simultaneously meet both energy and operating reserve requirements) such that the appropriate price signal for demand response and efficient incentives for new investment when required. During non-shortage conditions (*i.e.*, most of the time), the expectation of competitive market outcomes is no different in energy-only than in capacity markets.

Hence, in an energy-only market, it is the expectation of both the magnitude of the energy price during shortage conditions and the frequency of shortage conditions that will attract new investment when required. In other words, the higher the price during shortage conditions, the fewer shortage conditions that are required to provide the investment signal, and vice versa. While the magnitude of price expectations is determined by the PUCT energy-only market rules, it will remain an empirical question whether the frequency of shortage conditions over time will be optimal such that the market equilibrium produces results that satisfy the reliability planning requirements (*i.e.*, the maintenance of a minimum 12.5 percent planning reserve margin).

Finally, the PUCT's energy-only market rule provides that the IMM may conduct an annual review of the effectiveness of the SPM. The IMM anticipates performing such a review in 2008 that will focus on the results of the first year of operation under the new rules, the outlook for future years, and potential modifications, if any, that may be required to ensure that the energy-only market achieves its intended objectives.

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. Under the ERCOT Protocols, scheduled load must be balanced with scheduled

resources for each QSE for each settlement interval; however, there is no requirement that scheduled load be reflective of the actual load of a QSE. Additionally, QSEs may balance some or all of their scheduled load with resources scheduled from ERCOT. Because the financial effect of scheduling resources from ERCOT to balance a load schedule is the same as if the load were unscheduled, in this section, we adjust the load schedules by subtracting the amount that consists of resources scheduled from ERCOT.

To provide an overview of the scheduling patterns, Figure 32 shows a scatter diagram that plots the ratio of the final load schedules to the actual load level during 2006. The ratio shown in the figure will be greater than 100 percent when the final load schedule is greater than the actual load.

Figure 32: Ratio of Final Load Schedules to Actual Load
All ERCOT – 2006

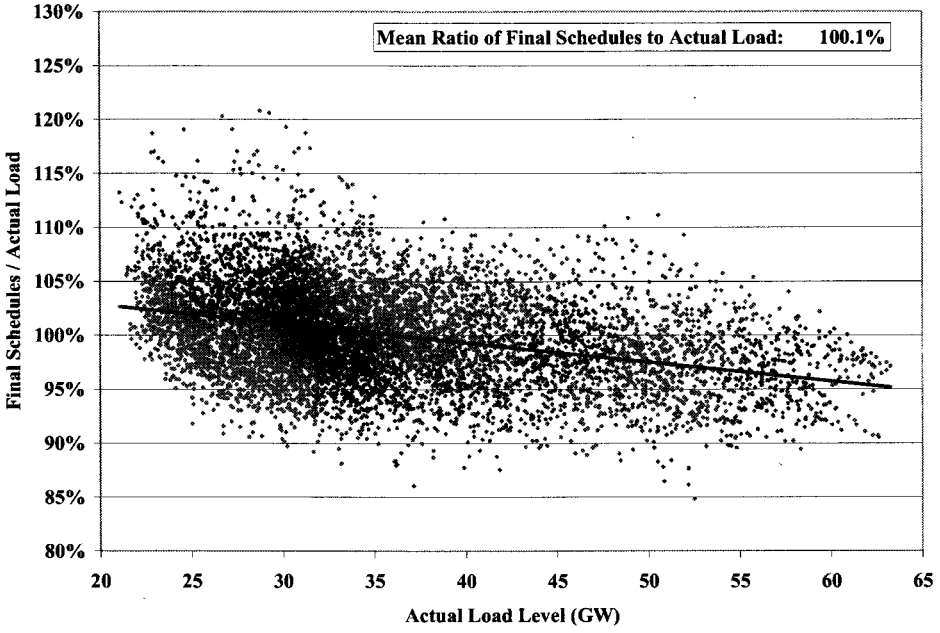


Figure 32 shows that final load schedules generally come very close to actual load in the aggregate, as indicated by an average ratio of the final load schedules to actual load of 100.1 percent. However, the figure also includes a trend line indicating that the ratio of final load schedules to actual load tends to decrease as load rises. In particular, the ratio given by the trend line is above 100 percent for loads under 37 GW and declines to 95 percent at higher load levels. The overall pattern shown in the figure above is similar to 2005, which exhibited the same

downward trend in final load schedules relative to actual load, although the average ratio was 101.2 percent.

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 a significant portion of 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. Third, some market participants may not have contracted for sufficient resources to cover their peak load and may, therefore, not be able to fully schedule their load.

Figure 33: Average Ratio of Final Load Schedules to Actual Load by Load Level All Zones – 2006

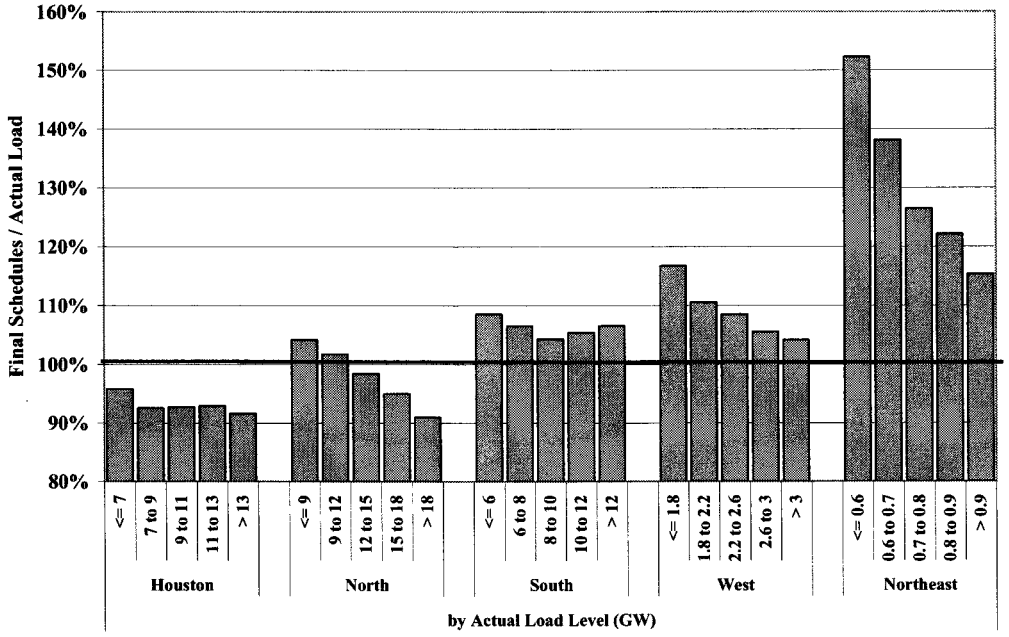


Figure 33 is a further analysis of final load schedules that shows the ratio of final load schedules to actual load evaluated at five different load levels for each of the ERCOT zones.

Figure 33 shows that:

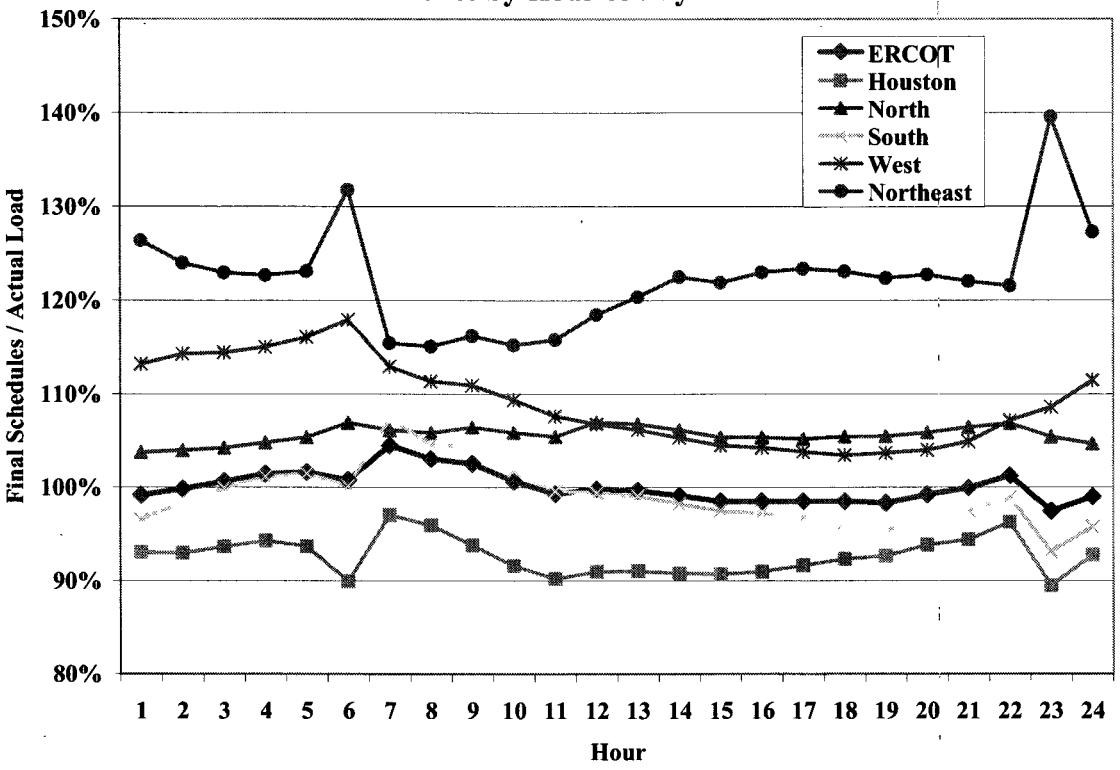
- The final schedule quantity decreases in each of the five zones as actual load increases, with the exception of the South zone which remains relatively flat as load changes.
- The South and West Zones are generally over-scheduled, although the ratios decline slightly as load increases.
- The Northeast 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 on average is about 190 MW.
- Houston is under-scheduled at most load levels, ranging from 4 percent at lower load levels up to 8 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. It should be noted that, regardless of the relationship between the aggregate scheduled load and actual load, individual QSEs may be significant net sellers or purchasers in the balancing energy market.

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.

To further analyze load scheduling, Figure 34 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 34: Average Ratio of Final Load Schedules to Actual Load All Zones by Hour of Day - 2006



This figure shows that on an ERCOT-wide basis, final schedules are close to actual load (between 99 percent and 102 percent) during hours ending 1 to 6. At hour ending 7, the ratio rises to 105 percent, the highest of any hour. By hour ending 10 through the remainder of the day, the ratio declines to a range between 97 percent and 101 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 34 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

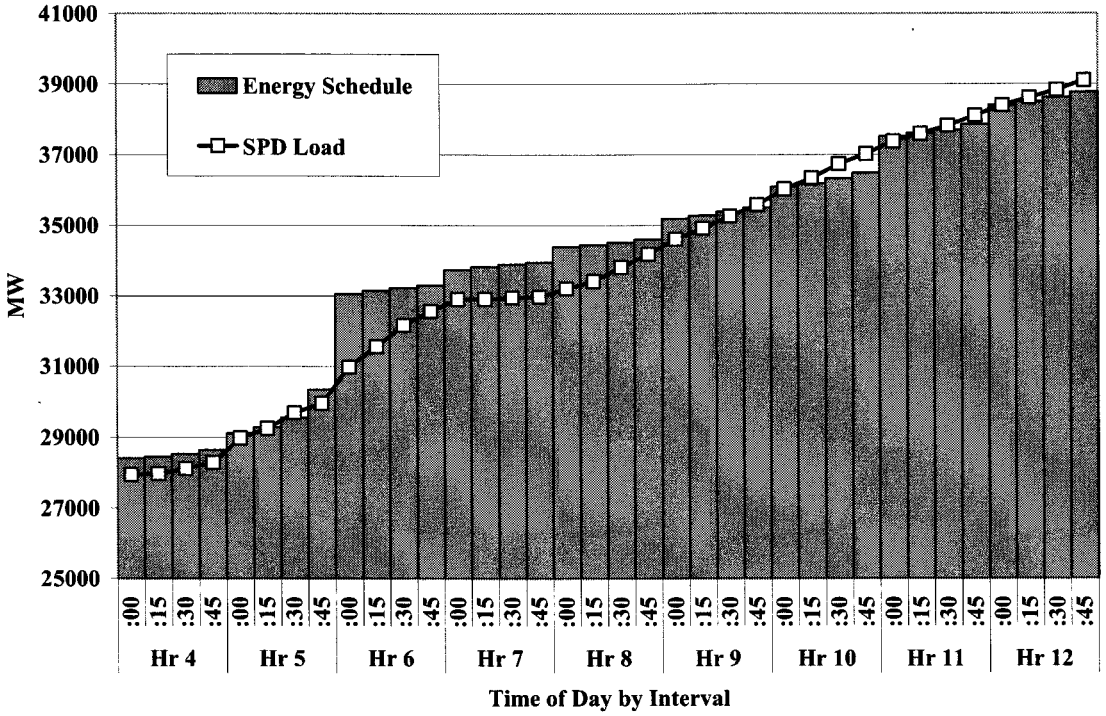
In the previous section, we analyzed balancing energy prices and load and found that while balancing energy prices are correlated to real-time load levels, other factors also have substantial

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 35 shows average energy schedules and actual load for each interval from 4 AM to 1 PM during 2006.

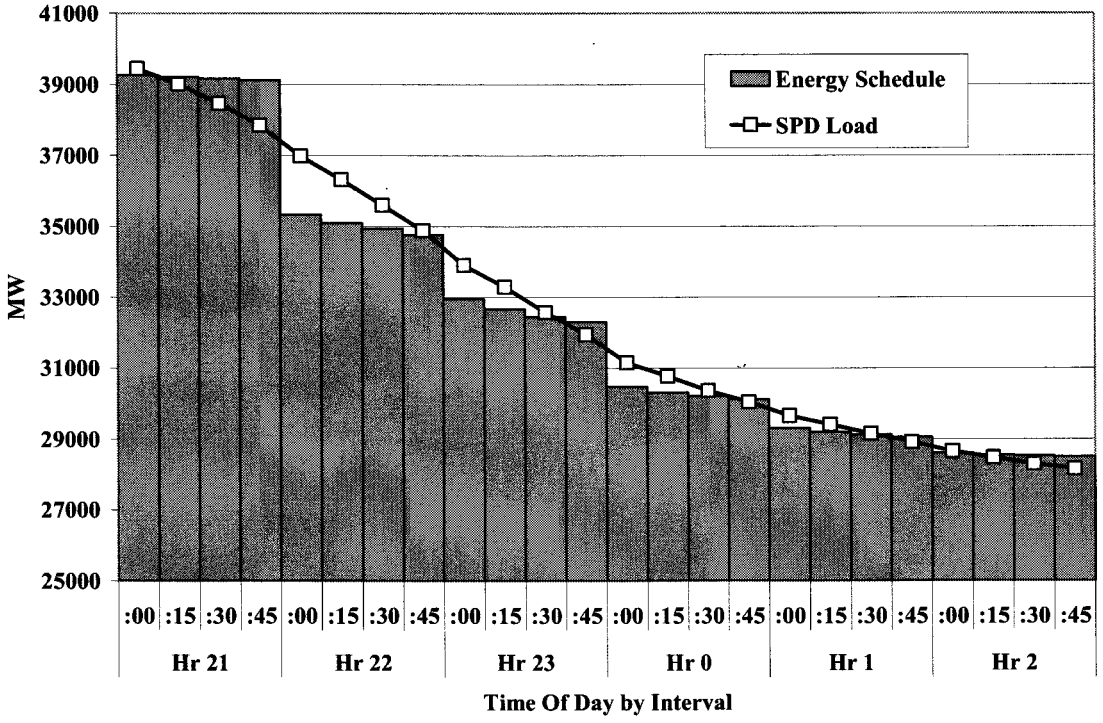
In general for ERCOT as a whole, energy schedules that are less than the actual load result in balancing energy purchases while energy schedules higher than actual load result in balancing energy sales. On average, load increases from approximately 28 GW to almost 39 GW in the nine hours shown in Figure 35. 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 35: Final Energy Schedules during Ramping-Up Hours 2006



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 over 2.7 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 36 shows average energy schedules and load for each interval from 9 PM to 3 AM during 2006.

Figure 36: Final Energy Schedules during Ramping-Down Hours 2006



On average, load drops from approximately 39 GW to less than 29 GW in the six hours shown in Figure 36. The average decrease per 15-minute interval is approximately 417 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-down 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 much of the generation in ERCOT is scheduled by QSEs that submit energy schedules that change hourly. 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.

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 35 and Figure 36). This analysis is similar to that shown in Figure 16 and Figure 17, except instead of showing balancing energy prices relative to load, we show balancing energy prices relative to balancing energy deployments. Figure 37 shows the analysis for the ramping-up hours.

Figure 37: Balancing Energy Prices and Volumes Ramping-Up Hours – 2006

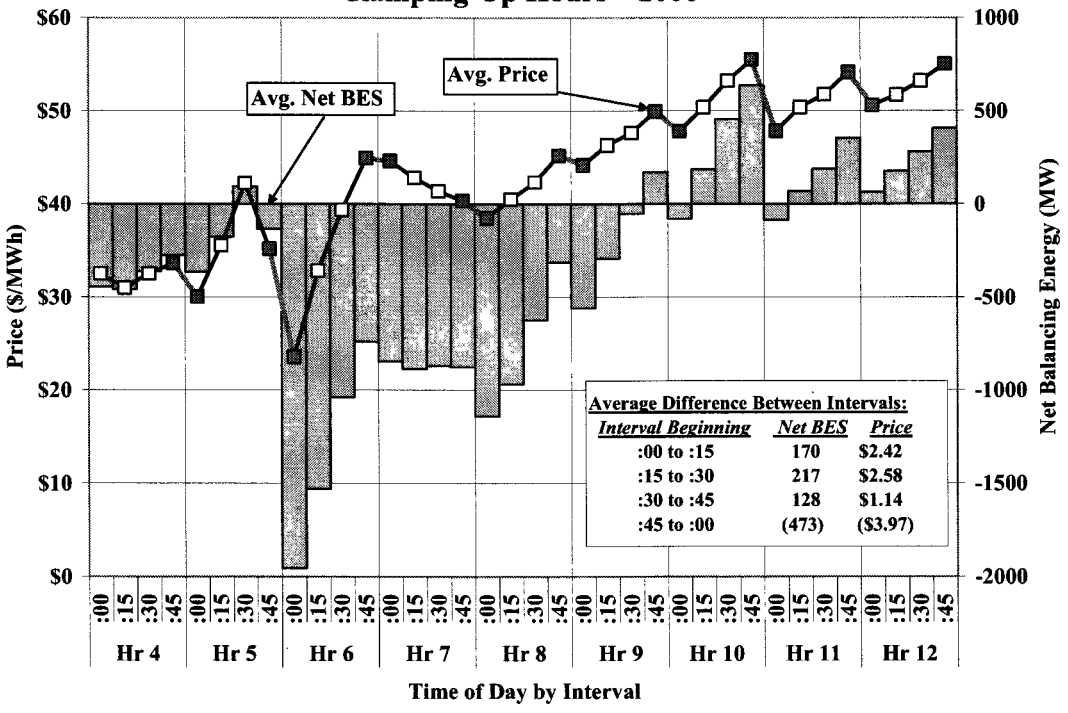
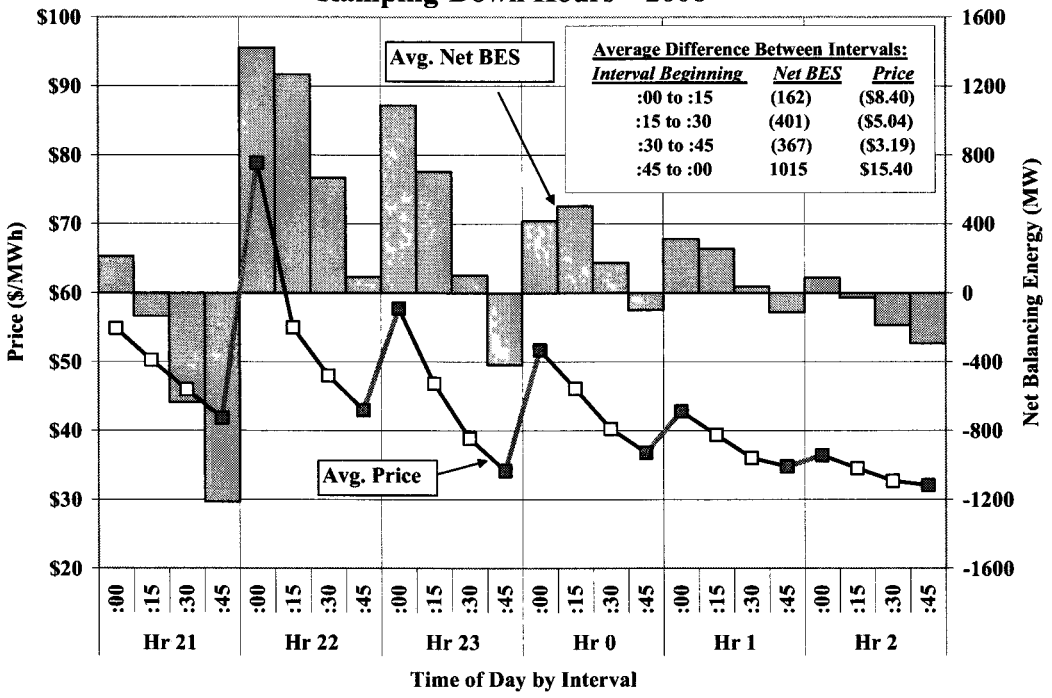


Figure 37 reveals two key aspects of the balancing energy market. First, as discussed above, balancing energy prices are highly correlated with balancing energy deployments. Second, with the exception of hour 7, there is a distinct pattern of increasing purchases during the hour. At the

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 38 shows the same analysis for the ramping-down hours. As discussed later in this section, most of these inefficiencies are due to structural issues that are inherent to the zonal market design, and implementation of the nodal market by 2009 will largely resolve these inefficiencies.

Figure 38: Balancing Energy Prices and Volumes
Ramping-Down Hours – 2006



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), many QSEs schedule only on an hourly basis, making little, or no changes on a 15-minute basis. It is

primarily the scheduling patterns by the QSEs that schedule on an hourly basis that result in the balancing energy deployments and prices shown in Figure 37 and Figure 38.

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 previous reports, and has continued to be a concern in 2006. To address this issue, we have previously 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. However, because of the resource demands and the timeframe for the nodal transition, such changes will not be accommodated in the zonal market design. 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.

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). There are 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. These issues were discussed in detail in the 2005 SOM Report.[18] The operational implications associated with these issues continued in 2006 and will likely continue until the current zonal market design is replaced. However, each of these issues will be significantly ameliorated or eliminated with the implementation of the nodal market.

D. Balancing Energy Market Offer Patterns

In this section, we evaluate balancing energy offer patterns by analyzing the rate at which capacity is offered. In Figure 39, we show the average amount of capacity offered to supply balancing up service relative to all available capacity. The analysis in this section differs from similar analyses in prior reports in the following important respect. In prior reports, un-offered capacity calculations included capacity that existed but was not offered. They did not attempt to quantify the amount of un-offered capacity that was actually available, and practicable to offer, given the ERCOT scheduling timelines, operating rules and conditions, and technical or commercial limitations that might limit a QSE's ability to offer capacity in the ERCOT market. In contrast, the approach used for the analysis of un-offered capacity in this section is focused on online, available capacity for which there is a reasonable expectation that the energy can be produced in light of the factors and considerations listed above. Specifically, the methodology for determining the quantities of un-offered capacity in this section are as follows:

Un-offered Capacity is equal to:

> Total Online Capacity plus qualified, off-line quick-start combustion turbines not providing non-spinning reserve;

[18] 2005 SOM Report at 68-76.

Less:

- Scheduled Generation;
- Up Balancing Energy Offers;
- Residual Reliability Must Run ("RMR") capacity;
- Residual Qualifying Facility ("QF") capacity from "non-bid" QF resources;
- Residual capacity from wind turbines;
- Scheduled energy (25 percent) from wind turbines that are not in wind-only QSEs;
- Generation Regulation Up obligation;
- Generation Responsive Reserve obligation; and
- Non-spinning Reserve obligation met by online resources.

The balancing energy offers are divided into that which is ramp-constrained, and would not actually be capable of supplying balancing energy in a single 15-minute interval, and that which is non-ramp-constrained, and thus would be available to supply balancing energy in a single 15-minute interval. Total capacity includes the maximum capacity of resources that are flagged as online in the final resource plan submitted by the QSE, as well as qualified, off-line quick start units that are not flagged as providing non-spinning reserve. Scheduled generation, regulation up, responsive reserve from generation resources and up balancing energy offers are deducted from the total capacity. Non-spinning reserve is deducted from the total online capacity for each QSE to the extent that the QSE has insufficient offline capability flagged as non-spinning reserve to meet it obligation. Residual RMR capacity is deducted from total capacity because, while such capacity could technically be offered, the financial incentives as set forth in the ERCOT Protocols are insufficient to provide a reasonable expectation that the residual RMR capacity would be offered. Capacity from a QF that is designated as "non-bid" is also deducted from the total online capacity. Under the ERCOT Protocols, QFs are allowed to specify capacity as "non-bid" for the purpose of local congestion management to reflect technical or commercial limitations associated with their specific operating requirements; therefore, such capacity is not reasonably expected to be offered as balancing energy. Residual wind capacity is deducted from the total online capacity to reflect the uncontrollable nature of wind turbines.

Finally, 25 percent of the scheduled wind generation from non-wind-only QSEs is deducted from total capacity to reflect the fact that, to the extent the wind does not produce as scheduled, the

portfolio balancing requirement for non-wind-only QSEs requires that sufficient capacity be reserved for this purpose. The final result of these deductions from the total online capacity plus qualified quick-start units that are not flagged as providing non-spinning reserve is the quantifiable un-offered capacity that could practicably and reasonably be expected to be offered, although, as discussed later in this section, there are several other structural impediments to offering even this capacity that are more difficult to quantify. The offered and quantifiable un-offered capacity data is shown for the peak hour of the day on a monthly average basis for 2006 in Figure 39.

Figure 39: Balancing Energy Offers Compared to Total Available Capacity Daily Peak Load Hours – 2006

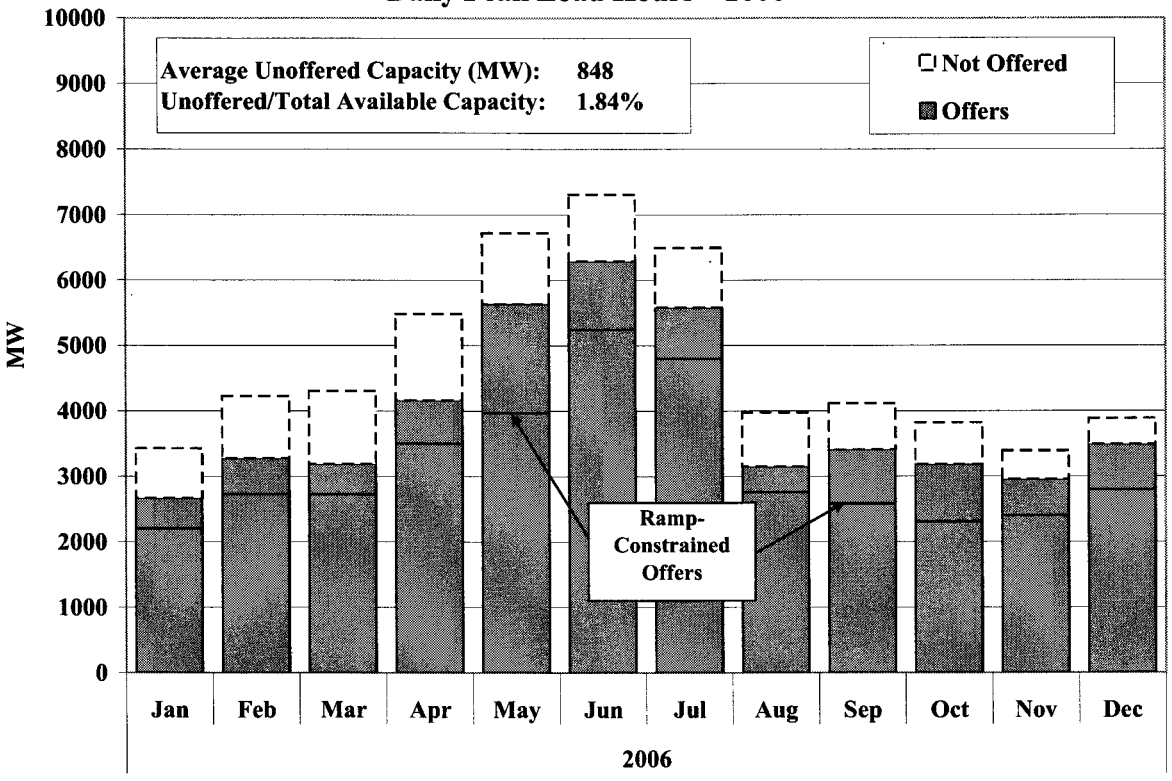
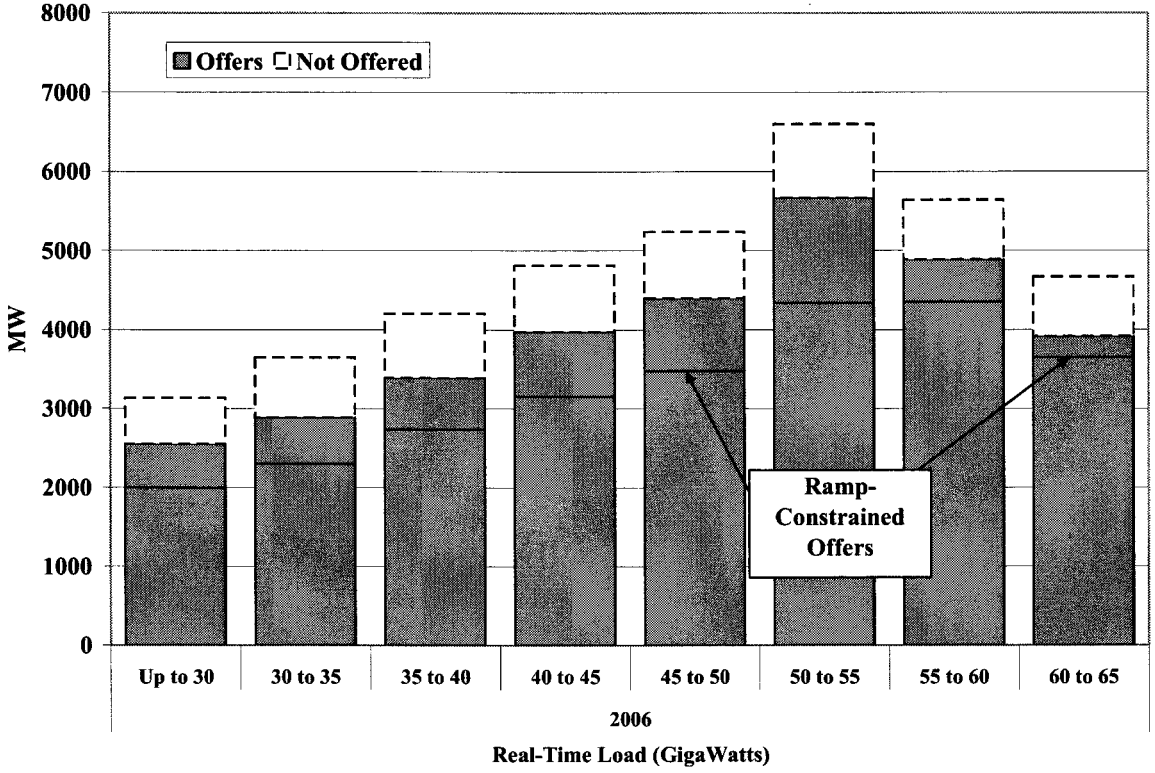


Figure 39 shows the trend in 2006 over time in quantities of energy available and offered to the balancing energy market. Up balancing offers are divided into the portion that is capable of being deployed in one interval and the portion which would take longer due to portfolio ramp rate offered by the QSE (*i.e.*, "Ramp-Constrained Offers").

Un-offered energy can raise competitive concerns to the extent that it reflects withholding by a dominant supplier that is attempting to exercise market power. To investigate whether this has

occurred, Figure 40 shows the same data as the previous figure, but arranged by load level for daily peak hours in 2006. Because prices are most sensitive to withholding under the tight conditions that occur when load is relatively high, increases in the un-offered capacity at high load levels would raise competitive concerns.

Figure 40: Balancing Energy Offers Compared to Total Available Capacity Daily Peak Load Hours – 2006



The figure indicates that in 2006, the average amount of capacity available to the balancing market increased gradually up to 55 GW of load and then declined at higher levels. The decline in balancing energy available at higher load levels is associated with the fact that scheduled generation increases at higher load levels, thereby leaving less residual capacity available to be offered as balancing energy. As indicated in the figure, the quantity of un-offered capacity does not change significantly as load levels increase.

The pattern of un-offered capacity shown in Figure 40 does not 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 do not change significantly as load levels increase. Based on this analysis and other analyses in the report at the supplier level, we do not find that the un-offered capacity raises potential competitive concerns.

In regard to the residual un-offered capacity shown in the previous two figures, there are several possible explanations for the quantity of un-offered on-line and quick start capacity that was not quantifiable in the preceding analysis. First, issues related to ramp rates can affect the offer levels. Currently, a QSE is able to submit one up-balancing ramp rate for its portfolio per hour per zone, and the ramp capability tends to decrease as more of the offer is deployed. Thus, many QSEs may feel compelled to not offer slow ramping capability 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 compel it to dispatch its high-cost resources at a loss.

Second, QSEs are subject to compliance measures in relation to their performance, which may include penalties. This may limit a QSE's willingness to adopt a very aggressive offer strategy in consideration of operational risks in real-time that may affect its ability to perform at the outer bounds of its rated capability. In aggregate, if such risks were managed by a conservative reduction of offered capacity of just one percent, this would represent 500 MW of un-offered capacity in an hour where 50,000 MW of rated generation capability was online.

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. Also, many generators cannot operate in the duct firing range and provide regulation simultaneously, so that dispatching in this range for energy could result in non-performance in the provision of ancillary services.