E. Resource Plan Changes

QSEs must have sufficient generation on-line to support their energy schedules and offers, and they are required to inform ERCOT about which resources they plan to use to satisfy their obligations. They do this by submitting resource plans at various points in the day-ahead and the operating day. While QSEs are expected to make their best effort to accurately forecast how they will operate their units, the resource plans are not financially binding and can be changed until shortly before real-time.[19] Resource plans are used by ERCOT in some of its reliability assessments before real-time and to make additional commitments to maintain reliability. Therefore, it is important for ERCOT to have accurate information in the resource plans that QSEs submit in order to avoid taking unnecessary and sometimes costly actions to maintain reliability.

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

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

We first analyze the behavior of suppliers that are the primary recipients of payments by ERCOT for out-of-merit capacity or replacement reserves. OOMC or RPRS occurs when ERCOT

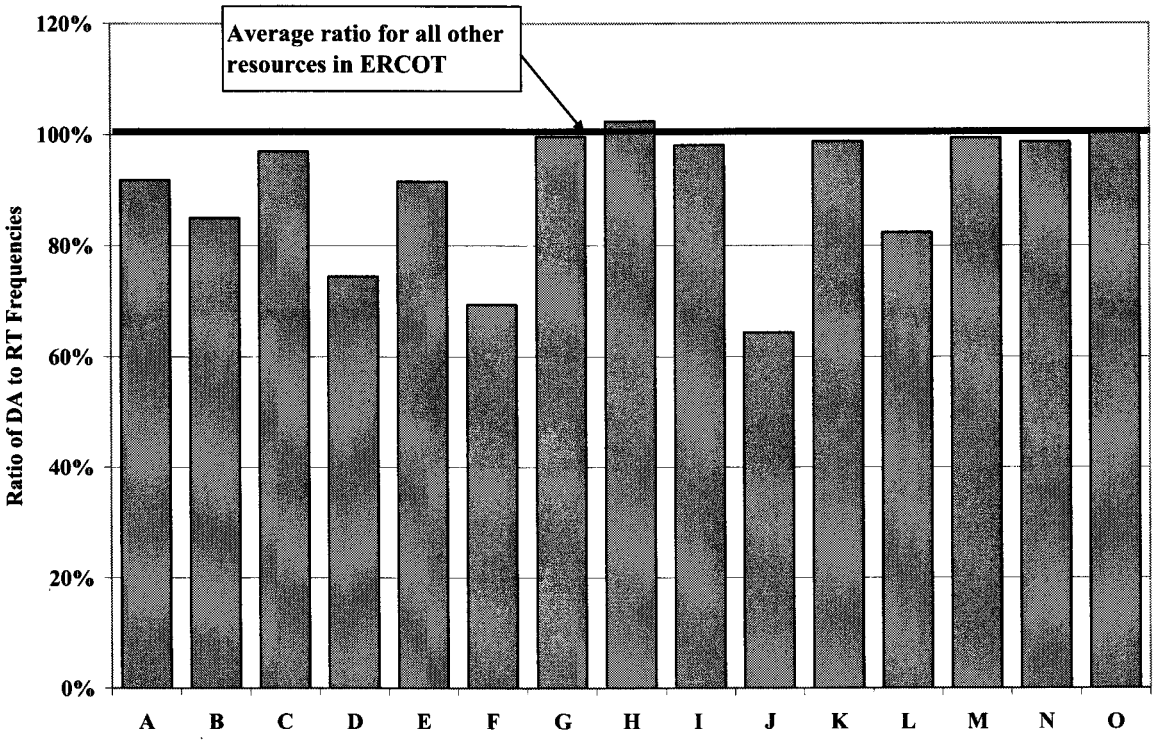
19 While resource plans are not financially binding, the real-time planned generation is used in the OOME payment formulas to determine the amount of megawatts deployed by the OOME instruction.

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

Because of the incentives presented by the OOMC and RPRS processes, we would expect suppliers that anticipate having units committed out-of-merit and that would benefit from the resulting payments to avoid showing the units as committed until after the out-of-merit commitments are announced. We examined the patterns of commitment for units that receive substantial OOMC and RPRS payments. Figure 41 shows the ratio of day-ahead resource plan commitments to actual real-time commitments during 2006 for the 15 generating plant sites receiving the largest OOMC and RPRS payments.[20] The generating plants in the figure are sequenced from highest to lowest payment from left to right, and the total payments for all generating plants in the figure constitute two-thirds of the total OOMC and RPRS payments in 2006. Hours when the resources are under OOMC, RPRS or OOME instructions are not included in order to assess systematic changes made voluntarily by market participants. The units are shown in decreasing order of payments received from ERCOT. To show how the commitment of these units compares to all other units in ERCOT, the figure also shows the capacity-weighted average ratio of day-ahead to real-time resource plan commitments for all units.

[20] For the purpose of this analysis, all generating units at the same electrical location are group into a single generating plant.

Figure 41: Ratio of Day-Ahead to Real-Time Resource Plan Commitments* Frequent OOMC Resources – 2006



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

While most of the generating plants shown in Figure 41 have ratios that are comparable to the market as a whole that reflects consistency between the day-ahead and real-time resource plans, a minority of the generating plants have ratios less than 80 percent. The results shown in this figure are consistent with the concern that some QSEs may wait until after the OOMC and RPRS process to commit units that are necessary for reliability.

For the generating plants shown in Figure 41, uplift payments for OOMC and RPRS commitments are substantial enough to provide significant incentives to behave in ways that maximize the likelihood of receiving them. Figure 41 suggests that some QSEs with resources that frequently receive OOMC or RPRS instructions may delay the decision to commit those units until after ERCOT determines which resources to select for OOMC or RPRS. This approach to address capacity insufficiency in the Protocols has several deleterious effects on the market. First, ERCOT incurs OOMC and RPRS costs to commit resources that are otherwise economic and that should be committed voluntarily without supplemental payments. Second,

when resources are committed out-of-merit, some other resources committed in day-ahead resource plans will no longer be economic. This can result in over-commitment of the system. However, the QSE generally has the opportunity to modify its other commitments after it receives the OOMC or RPRS instruction and often does so. Third, this conduct tends to undermine the accuracy of the information that ERCOT depends on to manage reliability. Ultimately, this can cause ERCOT to take a variety of costly actions, including making out-of-merit commitments that should not be necessary. These problems stem from the de-centralized process for unit commitment under the current market design, and underscore the reliability and efficiency benefits of the centralized commitment process that will be implemented with the nodal market re-design.

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

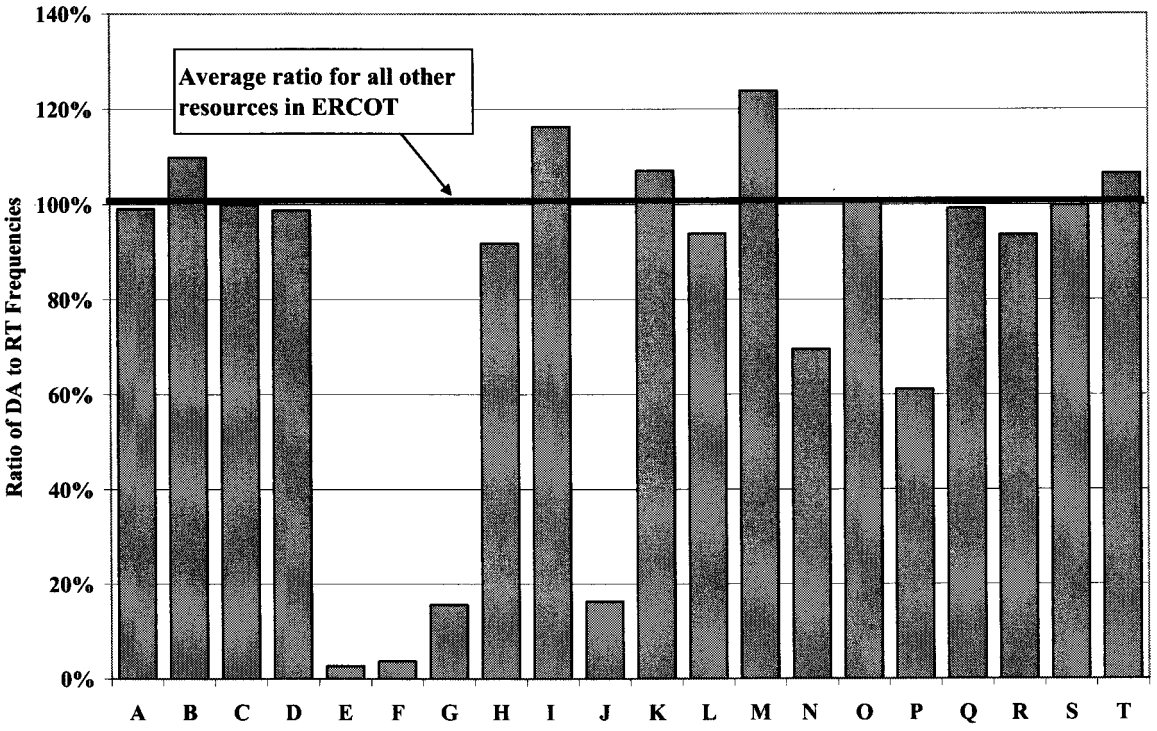
If a supplier is able to predict which of its units may be dispatched out-of-merit, it may under-schedule those units and over-schedule other units in its portfolio.[21] Although this resource plan output may not be efficient, it can be effective at compelling an OOME instruction and the associated uplift payment. Following the OOME instruction, the supplier can adjust its over-scheduled units to restore an economic dispatch pattern. If the supplier can accurately predict when the units will be called out-of-merit, this strategy can generate significant uplift payments. When the unit is not called for out of merit dispatch, the supplier can adjust the output levels of the units in its portfolio to correct the inefficient schedule.

Under this type of strategy, one would expect that units often needed to resolve congestion would be frequently under-scheduled. To test for this strategy, Figure 42 shows the ratio of real-

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

time resource plan scheduled output to actual generation for the 20 units that received the highest average payments for OOME Up per MWh of generation across all hours of 2006.[22]

Figure 42: Ratio of Real-Time Planned Generation to Actual Generation* Frequent OOME-Up Resources – 2006



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

To include only the scheduling and dispatch decisions made solely by the supplier, the ratio does not include hours when the resource was under OOMC or OOME instructions. The 20 resources shown in Figure 42 are presented in decreasing order of average payments, from $3.24 per MWh of generation across all hours for the unit on the far left to $0.22 per MWh for the unit on the far right. The generation-weighted average ratio of real-time resource plan output to actual generation for the whole ERCOT market is also shown for reference.

Of the 20 resources shown in Figure 42, 4 have ratios of less than 50 percent while ten have ratios between 50 and 100 percent. The other units in ERCOT had a weighted average ratio of

22 To focus on the most significant units, the analysis excludes resources received OOME up instruction less than 5 times within the year as well as resources that operated in-merit for fewer than 10 hours.

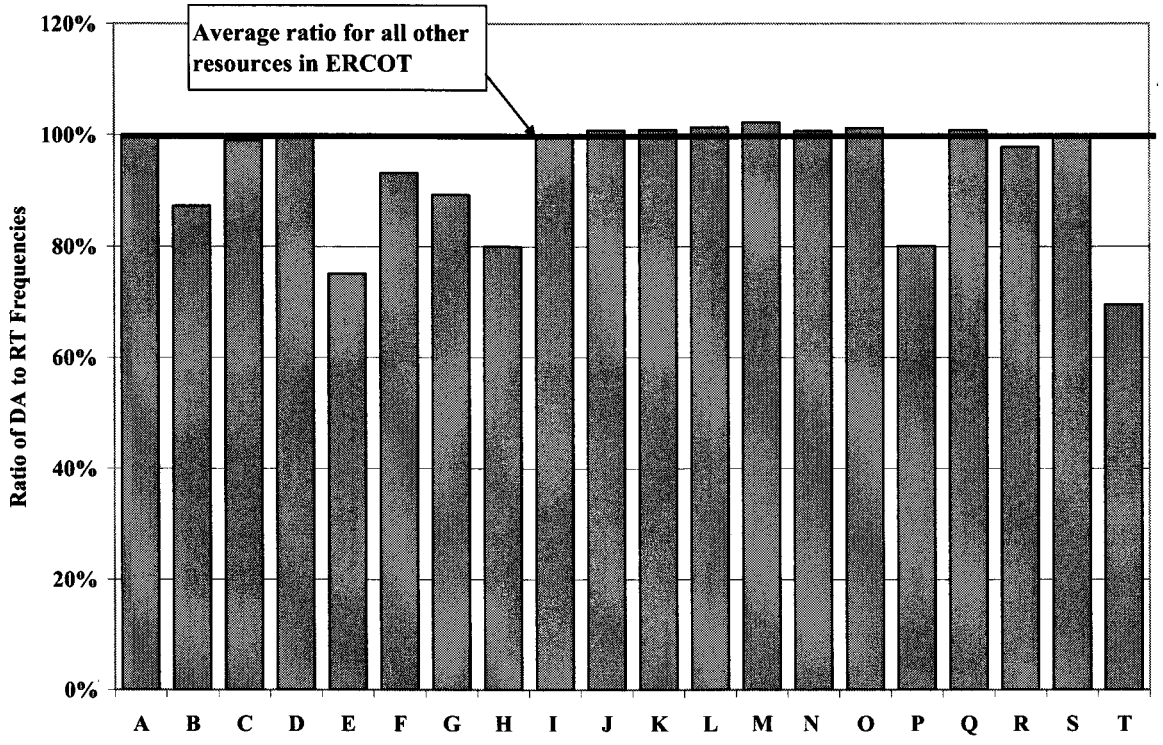
101 percent during the period, reflecting consistency between the scheduled output and actual generation. The data suggests that resources frequently providing OOME up are sometimes included by the QSEs in the real-time resource plans at output levels that are lower than their actual output. This is consistent with the hypothesis that the OOME procedures may provide inefficient incentives that lead QSEs to submit inaccurate resource plans.

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

Figure 43 shows the ratio of real-time resource plan output to actual generation for the twenty resources that earned the highest average payments for providing OOME down (on per MWh basis) in 2006.[23] The figure shows units that received the highest OOME down payments for their total production. The resources are shown in decreasing order of the average OOME down payments received per MWh of output, ranging from $0.38 per MWh on the far left to $0.04 per MWh on the far right. For comparison purposes, the figure also shows the generation-weighted average ratio of real-time resource plan output to actual generation for all other units.

[23] This analysis excludes resources with received OOME down instruction less than 5 times within the year.

Figure 43: Ratio of Real-Time Planned Generation to Actual Generation* Frequent OOME-Down Resources – 2006



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

None of the twenty resources shown in Figure 43 had a ratio that was significantly above 100 percent. The figure above reflects good consistency between the planned output level and actual generation for OOME down units. Thus, there is no indication that frequent OOME down units have systematically over-scheduled their resources to earn more OOME uplift.

III. DEMAND AND RESOURCE ADEQUACY

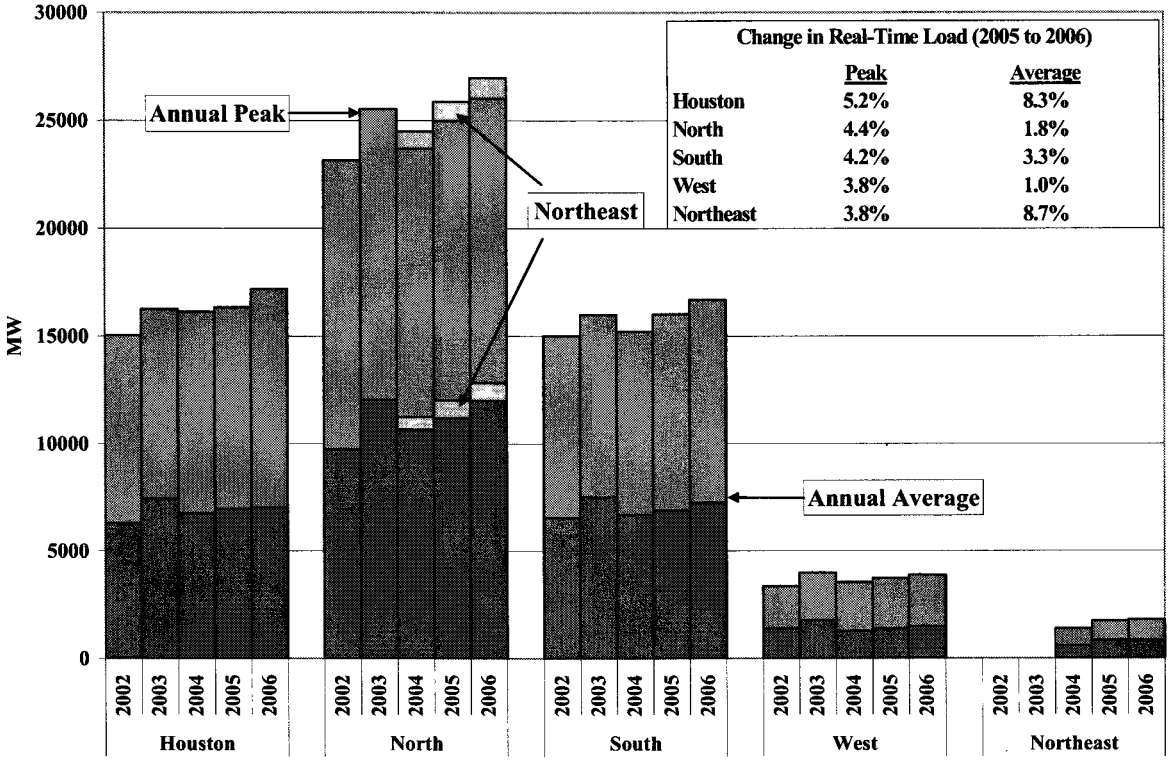
The prior sections of this report reviewed the market outcomes and provided analyses of a variety of factors that have influenced the market outcomes. This section reviews and analyzes the load patterns during 2006 and the existing generating capacity available to satisfy the load and operating reserve requirements.

A. ERCOT Loads in 2006

There are two important dimensions of load that should be evaluated separately. First, the changes in overall load levels from year to year can be shown by tracking the changes in average load levels. This metric will tend to capture changes in load over a large portion of the hours during the year. Second, it is important to separately evaluate the changes in the load during the highest-demand hours of the year. Significant changes in these peak demand levels have historically been very important and played a major role in assessing the need for new resources. The expectation in a regulated environment was that adequate resources would be acquired to serve all firm load, and this expectation remains in the competitive market. The expectation of resource adequacy is based on the value of electric service to customers and the damage and inconvenience to customers that can result from interruptions to that service. Additionally, significant changes in peak demand levels affect the probability and frequency of shortage conditions (*i.e.*, conditions where firm load is served but the maintenance of required operating reserves is challenged). Hence, both of these dimensions of load during 2006 are examined in this subsection and summarized in Figure 44.

This figure shows peak load and average load in each of the ERCOT zones from 2002 to 2006. It indicates that in each zone, as in most electrical systems, peak demand significantly exceeds average demand. The North Zone is the largest zone (about 37 percent of the total ERCOT load); the South and Houston Zones are comparable (with about 26 percent and 28 percent, respectively) while the West Zone and Northeast Zone are the smallest (with about 7 percent and 3 percent of the total ERCOT load). Figure 44 shows the annual non-coincident peak load for each zone. This is the highest load that occurred in a particular zone for one hour during the year; however, the peak can occur in different hours for different zones. As a result, the sum of the non-coincident peaks for the five zones was greater than the annual ERCOT peak load.

Figure 44: Annual Load Statistics by Zone*
2002 to 2006



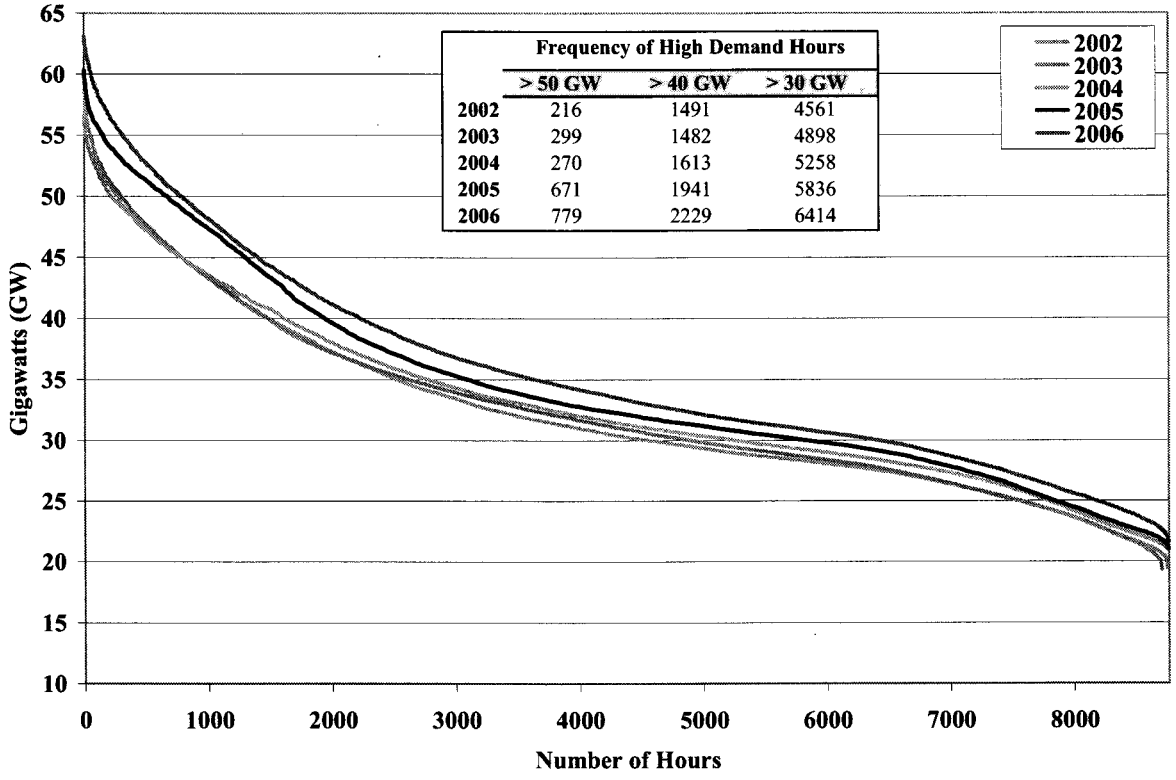
* The figure above is based on the load that SPD uses to schedule supply in the balancing energy market. This can differ from actual load in individual intervals.

No load statistics are shown for the Northeast Zone before 2004 because it was separated from the North Zone at the beginning of 2004. For comparison purposes, the Northeast Zone is also shown stacked with the North Zone from 2004 to 2006.

To provide a more detailed analysis of load at the hourly level, Figure 45 compares load duration curves for each year from 2002 to 2006. A load duration curve shows the number of hours (shown on the x-axis) that load exceeds a particular level (shown on the y-axis). ERCOT has a fairly smooth load duration curve, typical of most electricity markets, as most hours exhibit low to moderate electricity demand, with peak demand usually occurring during the afternoon and early evening hours of days with exceptionally high temperatures. The highest load hours occur in the summer months, and ERCOT dispatched generation to meet a record peak demand of 63 GW in August 2006.[24]

[24] This value is the total load to be served in real-time as represented in ERCOT's Scheduling, Pricing and Dispatch software (including transmission and distribution losses), and may differ from settlement values.

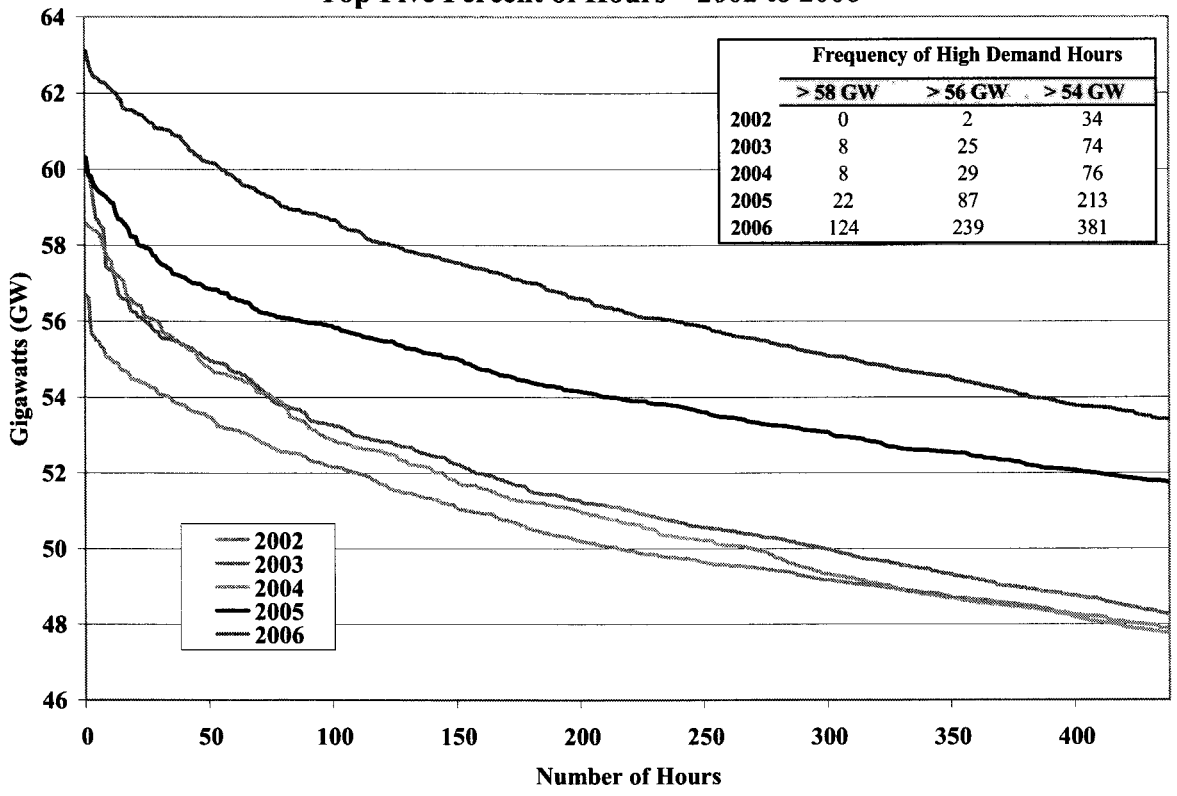
Figure 45: ERCOT Load Duration Curve
All Hours – 2002 to 2006



As shown in Figure 45 , the load duration curve for 2006 lies above the curves for the previous four years. Load increased more from 2005 to 2006 than it did in the previous three years on average. In 2006, there were 15 percent more hours when load exceeded 40 GW than in 2005.

To better show the differences in the highest-demand periods between years, Figure 51 shows the load duration curve for the five percent of hours with the highest loads. It shows that while load increased in each year from 2002 to 2005, the increase from 2005 to 2006 was much larger during the peak hours. Load exceeded 58 GW in 124 hours in 2006, 22 hours in 2005 and eight hours in 2003 and 2004. In 2002, demand was not higher than 58 GW in any hour. The same pattern prevailed at lower load levels with 2006 demand being considerably higher than in previous years.

Figure 46: ERCOT Load Duration Curve
Top Five Percent of Hours – 2002 to 2006



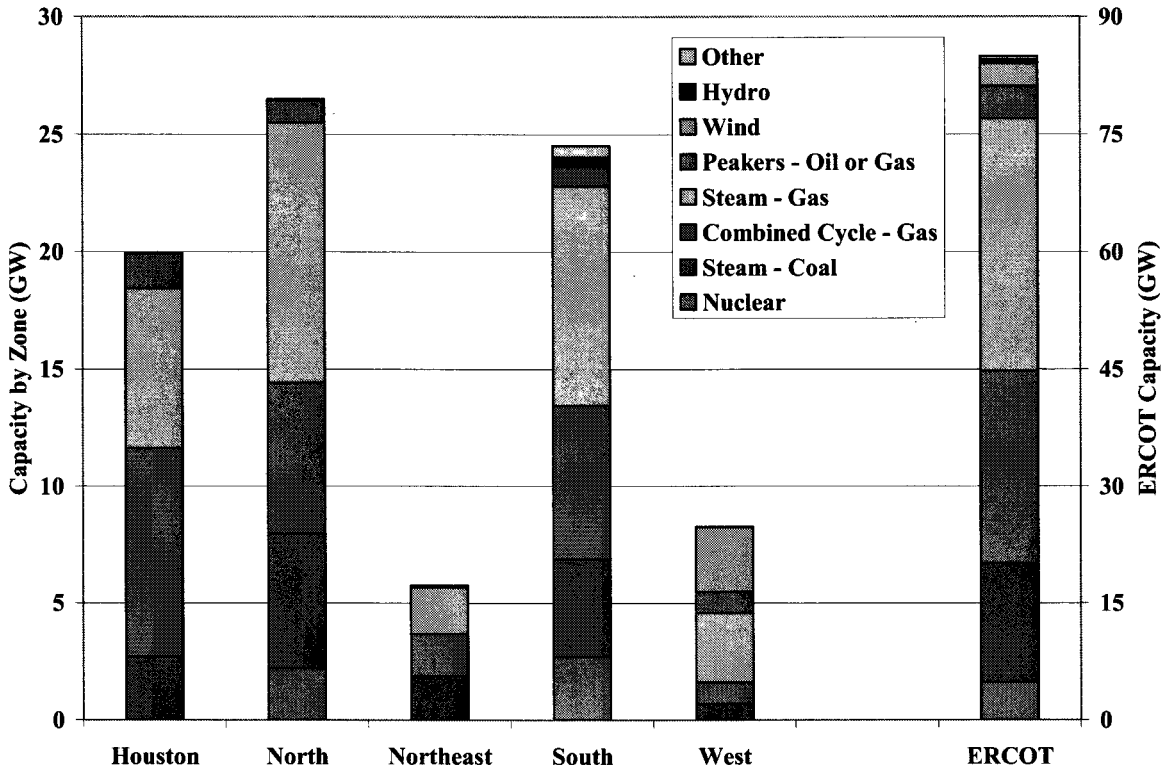
	Frequency of High Demand Hours		
	> 58 GW	> 56 GW	> 54 GW
2002	0	2	34
2003	8	25	74
2004	8	29	76
2005	22	87	213
2006	124	239	381

This figure also shows that the peak load in each year was roughly 15 to 25 percent greater than the load at the 95th percentile of hourly load. For instance, in 2006, the peak load value was over 63 GW while the 95th percentile was lower than 54 GW. This is typical of, and even somewhat flatter than, the load patterns in most electricity markets. This implies that a substantial amount of capacity, more than 9 GW, is needed to supply energy in less than 5 percent of the hours. This serves to emphasize the importance of efficient pricing during peak demand conditions to send accurate economic signals for the investment in and retention of these resources.

B. Generation Capacity in ERCOT

In this section we evaluate the generation mix in ERCOT. With the exception of the wind resources in the West Zone and the nuclear resources in the North and South Zones, the mix of generating capacity is relatively uniform in ERCOT. Figure 47 shows the installed generating capacity by type in each of the ERCOT zones.

Figure 47: Installed Capacity by Technology for each Zone 2006



The nuclear capacity is located in both the North and South Zones, and lignite and coal generation is also a significant contributor in ERCOT. However, the primary fuel in all five zones is natural gas (or sometimes oil) -- accounting for 76 percent of generation capacity in ERCOT as a whole, and 86 percent in the Houston Zone. Much of this natural gas-fired capacity represents relatively new combined-cycle units than have been installed throughout ERCOT over the past decade. These new installations have resulted in a small increase in the gas-fired share of installed capacity but have not changed the overall mix significantly, since the generators that have gone out of service during this period were primarily gas-fired steam turbines.

While ERCOT has coal/lignite and nuclear plants that operate primarily as base load units, its reliance on natural gas resources makes it vulnerable to natural gas price spikes. There is approximately 20,000 MW of coal and nuclear generation in ERCOT. Because there are very few hours when ERCOT load drops as low as 20,000 MW, natural gas resources will be dispatched and set the balancing energy spot price in most hours. Hence, although coal-fired and

///

nuclear units produce approximately half of the energy in ERCOT, they play a much less significant role in setting spot electricity prices.

The distribution of capacity among the ERCOT zones is similar to the distribution of demand. This is consistent with the legacy of investment under the regulated vertically integrated utilities when load and resources were largely integrated within separate control areas. The North Zone accounts for 31 percent of capacity, the South Zone 29 percent, the Houston Zone 23 percent, the West Zone 10 percent, and the Northeast Zone 7 percent. The North Zone and Houston are typically importers of power, while the Northeast Zone exports significant quantities because it has over two times more generation than its peak zonal load. Because large amounts of power flow out of the South Zone into the North Zone and Houston, the South-to-North CSC and the South-to-Houston CSC experienced the greatest amounts of congestion during 2006.

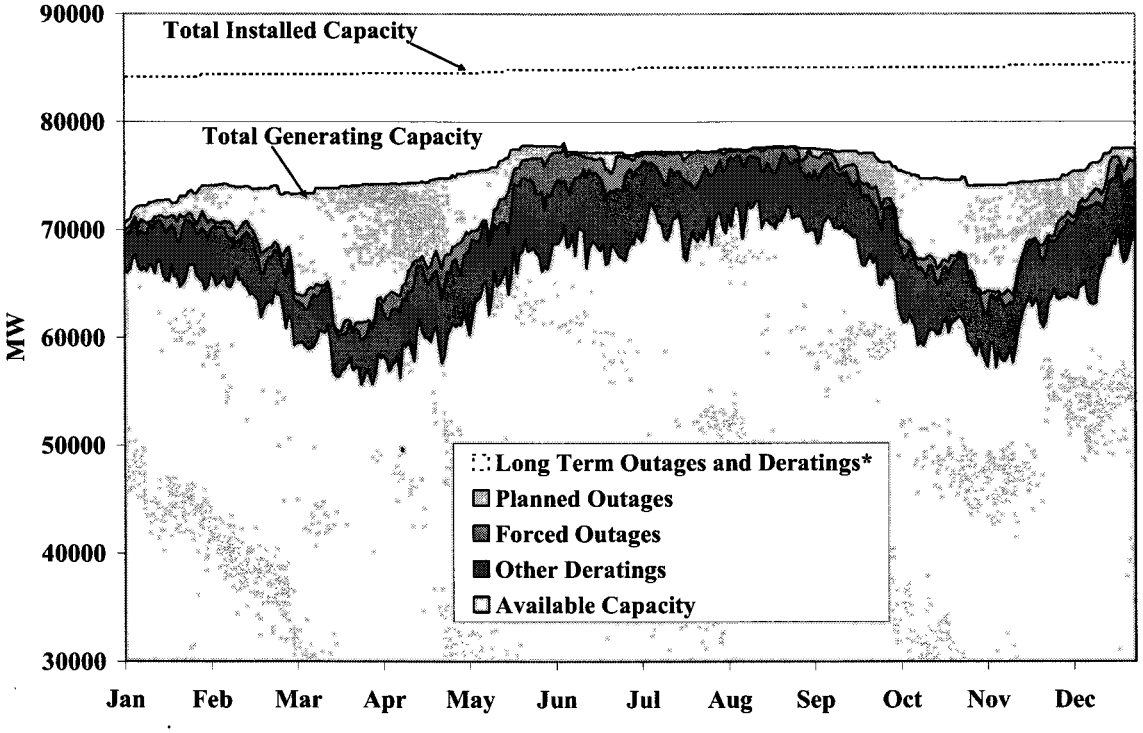
1. Generation Outages and Deratings

Figure 47 in the prior subsection shows that installed capacity far exceeds the annual peak load plus ancillary services requirements in ERCOT. This might suggest that the adequacy of resources is not a concern in ERCOT in the near-term, although resource adequacy must be evaluated in light of the resources that are actually available on a daily basis to satisfy the energy and operating reserve requirements in ERCOT. A substantial portion of the installed capability is frequently unavailable due to generator deratings. A derating is the difference between the maximum installed capability of a generating resource and its actual capability (or "rating") in a given hour. Generators can be fully derated (rating equals 0) due to a forced or planned outage. However, it is very common for generators to be partially derated (*e.g.*, by 5 to 10 percent) because the resource cannot achieve its installed capability level due to technical factors or environmental factors (*e.g.*, ambient temperature conditions).

In this subsection, we evaluate long-term and short-term deratings to inform our evaluation of ERCOT capacity levels. Figure 48 below shows a breakdown of total installed capability for ERCOT on a daily basis during 2006. This analysis includes all in-service and switchable capacity. The capacity in this analysis is separated into five categories: (a) long-term outages and deratings, (b) short-term planned outages, (c) short-term forced outages, (d) other short-term deratings, and (e) available and in-service capability.

The long-term deratings category includes any outages and deratings lasting for 60 days or longer while the remaining outages and deratings are included in the short-term categories. We generally separate the long-term outages because it provides an indication of the generating capacity that is generally not available to the market, which typically exceeds 10 GW. Long-term deratings can occur for several reasons. First, some of this capacity may be out-of-service for extended periods due to maintenance requirements. Second, if their owners predict that wholesale market prices will not be sufficiently high to justify the periodic costs required to keep them available, some units may go out-of-service temporarily. Third, the owners of some cogeneration plants routinely use steam output to support their processes rather than generate electricity. However, a large share of these deratings reflect output ranges on generating units that are not capable of producing up to the full installed capability level.

Figure 48: Short and Long-Term Deratings of Installed Capability**
2006



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

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

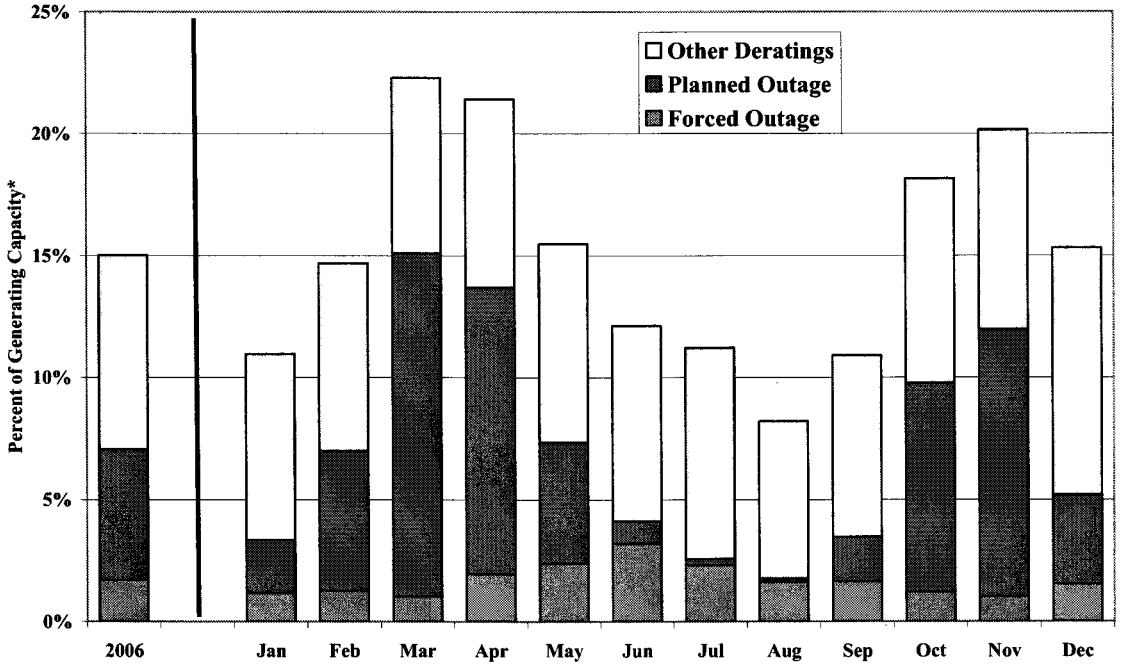
Figure 48 shows that installed capacity, including mothballed and switchable capacity, rose from 84 GW at the beginning of 2006 to 85 GW at the end of 2006. This increase is due to several

new generators coming on-line although it was diminished by several retirements. There were 1,028 MWs of new wind capacity coming on line from May through December 2006. The figure shows that the long-term outages and deratings fluctuated between 6 GW and 13 GW. The long-term outages and deratings also include over 8 GW of mothballed capacity.[25] These classes of capacity can be made available if market conditions become tighter as load rises.

As expected, short-term planned outages are relatively large in the spring and fall, decreasing to close to zero during the summer. Available in-service capacity fluctuated between 51 GW in March and 69 GW in August. The peak hour for the year required just over 63 GW to satisfy ERCOT's energy requirements plus approximately 4 GW for operating reserves and regulation-up requirements, resulting in surplus capacity of approximately 2 GW on that day.

The next analysis focuses specifically on the short-term forced outages and other short-term deratings. Figure 49 shows the average magnitude of the outages and deratings lasting less than 60 days for the year and for each month during 2006.

Figure 49: Short-Term Outages and Deratings*
2006



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

25 See "Report on the Capacity, Demand, and Reserves in the ERCOT Region," June 2006.

Figure 49 shows that total short-term deratings and outages were as large as 23 percent of installed capacity in the spring and fall, and dropped below 15 percent for the summer. Most of this fluctuation was due to anticipated planned outages, which ranged as high as 10 to 14 percent of installed capacity during March, April, October, and November. Short-term forced outages occurred more randomly, as would be expected, ranging between 1 percent and 4 percent of total capacity on a monthly average basis during 2006. These rates are relatively low in comparison to other operating markets, which can be attributed to a number of factors mentioned below.

First, these outages include only full outages (*i.e.*, where the resource's rating equals zero). In contrast, an equivalent forced outage rate is frequently reported for other markets, which includes both full and partial outages. Hence, the forced outage rate shown in Figure 49 can be expected to be lower than equivalent forced outage rates of other markets. Second, we were not confident that the forced outage logs received from ERCOT included all forced outages that actually occurred.

The largest category of short-term deratings was the "other deratings", which occur for a variety of reasons. The other deratings would include any short-term forced or planned outage that was not reported or correctly logged by ERCOT. This category also includes deratings due to ambient temperature conditions, cogeneration uses, and other factors described above. Furthermore, suppliers may delay maintenance on components such as boiler tubes, resulting in reduced capability. Because these deratings can fluctuate day to day or seasonally, some of the deratings are included in the "long-term outages and deratings" category while the others are included in this category. The other deratings were approximately 7 percent on average during the summer in 2006 and as high as 12 percent in other months. In conclusion, the patterns of outages do not indicate physical withholding or raise other competitive concerns. However, this issue is analyzed in more detail in Section V of this report.

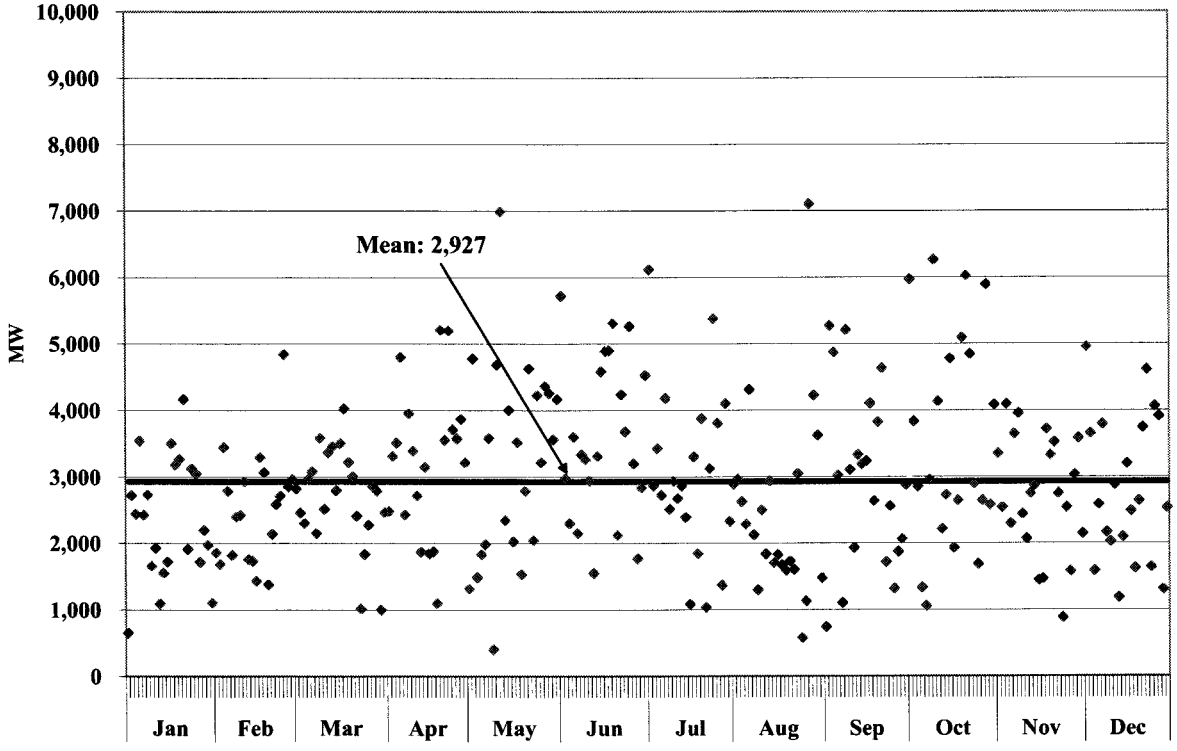
2. Daily Generator Commitments

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

This subsection evaluates the commitment patterns in ERCOT by examining the levels of excess capacity. Excess capacity is defined as the total online capacity plus quick-start[26] units minus the demand for energy, responsive reserve, up regulation and non-spinning reserve provided from online capacity or quick-start units. If the goal were to have no excess capacity, ERCOT would have to dispatch quick-start resources each day to meet its energy demand. Normally, however, because it is uneconomic to dispatch quick-start units for energy on most days, additional slow-starting resources with lower production costs are committed instead.

To evaluate the commitment of resources in ERCOT, Figure 50 plots the excess capacity in ERCOT during 2006.The figure shows the excess capacity in only the peak hour of each weekday because largest amount of additional generation commitment usually occurs at the peak hour. Hence, one would expect larger quantities of excess capacity in other hours.

Figure 50: Excess On-Line and Quick Start Capacity During Daily Peaks on Weekdays -- 2006



26 For the purposes of this analysis, "quick-start" includes simple cycle gas turbines that qualified to provide balancing energy.

Figure 50 shows that the excess on-line capacity during daily peak hours on weekdays averaged 2,927 MW in 2006, which is approximately 8 percent of the average load in ERCOT. This is a significant decrease from the average of 4,313 MW in 2005 and 6,627 MW in 2004. These decreases can be attributed in part to the continued increase in ERCOT load with a relatively static available supply, fewer quick-start gas turbines that were qualified to provide balancing energy, and a continuation of the trend from previous years of ERCOT committing fewer units via OOMC instructions and RMR.

The overall trend in excess on-line capacity also indicates a movement toward more efficient unit commitment across the ERCOT market; however, the current market structure is still based primarily upon a decentralized unit commitment process whereby each participant makes independent generator commitment decisions that are not likely to be optimal. Further contributing to the suboptimal results of the current unit commitment process is that the decentralized unit commitment is comprised of non-binding resource plans that form the basis for ERCOT's day-ahead planning decisions. However, these non-binding plans can be modified by market participants after ERCOT's day ahead planning process has concluded causing ERCOT to take additional actions that may be more costly and less efficient. Hence, the introduction of a day-ahead energy market with centralized Security Constrained Unit Commitment ("SCUC") that is financially binding under the nodal market design planned for implementation by 2009 promises substantial efficiency improvements in the commitment of generating resources.

C. Demand Response Capability

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

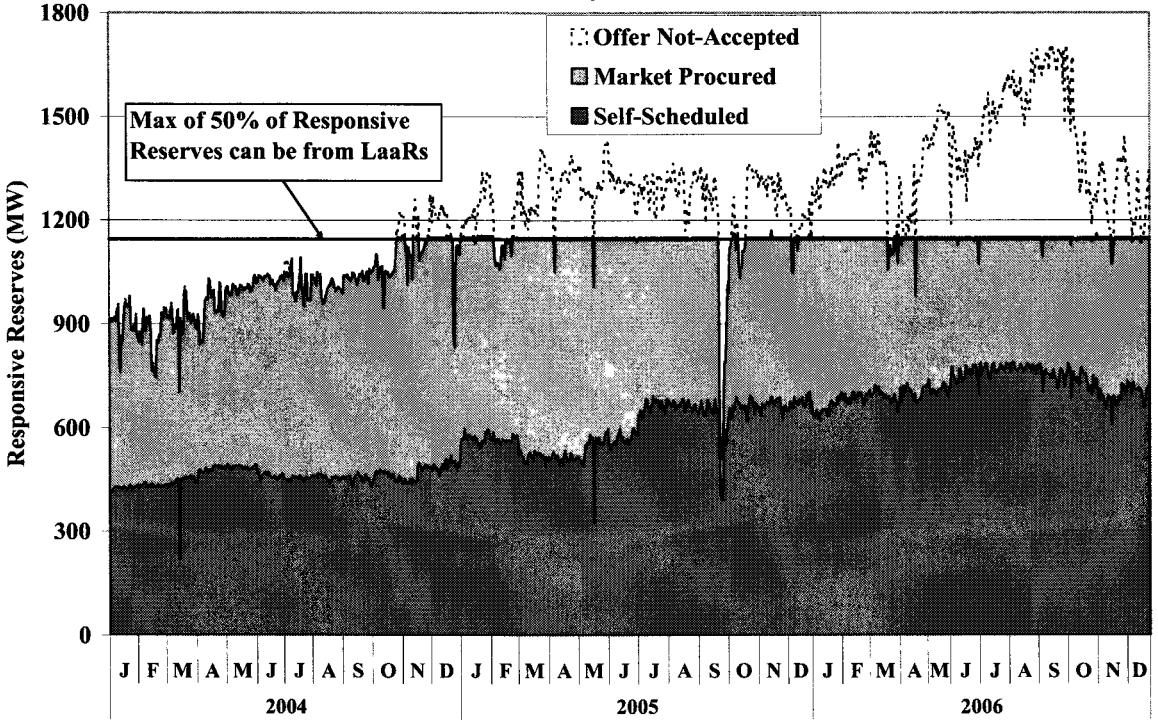
ERCOT allows qualified LaaRs to offer responsive reserves and non-spinning reserves into the day-ahead ancillary services markets. Qualified LaaRs can also offer blocks of energy in the

balancing energy market. LaaRs providing up balancing energy must have telemetry and must be capable of responding to ERCOT energy dispatch instructions in a manner comparable to generation resources. Those providing responsive reserves must have high set under-frequency relay ("UFR") equipment. A load with UFR equipment is automatically tripped when the frequency falls below 59.7 Hz.

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

As of December 2006, 1,985 MW of capability were qualified as LaaRs. These resources regularly provided reserves in the responsive reserves market, but never participated in the balancing energy market and only a very small portion participated in the non-spinning reserves market. There were no BULs registered with ERCOT in 2006. Figure 51 shows the amount of responsive reserves provided from LaaRs on a daily basis in 2006.

Figure 51: Provision of Responsive Reserves by LaaRs
Daily Average – 2006



The high level of participation by demand response sets ERCOT apart from other operating electricity markets. Figure 51 shows that the amount of responsive reserves provided by LaaRs gradually increased from about 900 MW at the beginning of 2004 to an average of 1,147 MW in 2006. The majority of this increase was procured through self-provision and bilateral agreements rather than the ERCOT administered auction. Currently, LaaRs are permitted to supply up to 1,150 MW of the responsive reserves requirement. In 2005 and 2006, it became commonplace for the 1,150 MW restriction to limit the set of demand resources that could provide responsive reserves. This has highlighted a flaw with the way that the ancillary services auction selects demand resources to provide responsive reserves.

The auction ranks responsive reserves providers according to their offer price from lowest to highest.[27] The auction goes up the offer stack until it reaches the 2,300 MW required quantity of reserves. However, if the auction reaches the 1,150 MW limit before meeting the 2,300 MW requirement, the offers of any additional LaaRs cannot be used and are discarded. In such cases, the marginal generator resource sets the clearing price for responsive reserves at a level that exceeds the offer prices of some of the unaccepted offers from LaaRs.

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

---

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

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

To improve the efficiency of responsive reserves pricing and incentives for suppliers, we recommend that ERCOT set separate prices for the two types of responsive reserves. The best way to accomplish this would be by having two responsive reserves constraints in the ancillary services auction: (i) that the responsive reserves procurement (including bilateral schedules) be greater than or equal to 2,300 MW and (ii) that the responsive reserves procurement from LaaRs (including bilateral schedules) be less than or equal to 1,150 MW. The clearing price paid to generators would be equal to the shadow price of the first constraint only, while the clearing price paid to LaaRs would be equal to the shadow price of the first constraint minus the shadow price of the second constraint.

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

ERCOT stakeholders considered this change in 2006 and, due to resource constraints, decided not to implement it in the current market and instead drafted a protocol revision to implement it in the nodal market. However, this protocol revision failed to receive the necessary two-thirds vote at the ERCOT Technical Advisory Committee in 2007; thus, there is currently no plan to

implement any of the changes described above for the RRS market. As previously discussed, the current mechanism for selecting providers and determining clearing prices for responsive reserves is inefficient and leads to excessive reliability costs for consumers. Therefore, we recommend that these changes be reconsidered for implementation in the nodal market design.

Although LaaRs are active participants in the responsive reserves market, they did not offer into the balancing energy or regulation services markets and their participation in the non-spinning reserves market averaged only 14 MW in 2006. This is not surprising because the value of curtailed load tends to be very high, and providing responsive reserves offers substantial revenue with very little probability of being deployed. In contrast, providing non-spinning reserves introduces a much higher probability of being curtailed. Participation in the regulation services market requires technical abilities that most LaaRs cannot meet at this point. Finally, prices in the balancing energy market have not been high enough to attract active load participation in that market. Hence, most LaaRs will have a strong preference for providing responsive reserves over regulation services, non-spinning reserves, or balancing energy.

IV. TRANSMISSION AND CONGESTION

One of the most important functions of any electricity market is to manage the flows of power over the transmission network by limiting additional power flows over transmission facilities when they reach their operating limits. In ERCOT, constraints on the transmission network are managed in two ways. First, ERCOT is made up of zones with the constraints between the zones managed through the balancing energy market. The balancing energy market model increases energy production in one zone and reduces it in another zone to manage the flows between the two zones when the interface constraint is binding, *i.e.*, when there is interzonal congestion. Second, all other constraints not defined as zonal constraints (*i.e.*, local congestion) are managed through the redispatch of individual generating resources. In this section of the report, we evaluate the ERCOT transmission system usage and analyze the costs and frequency of transmission congestion.

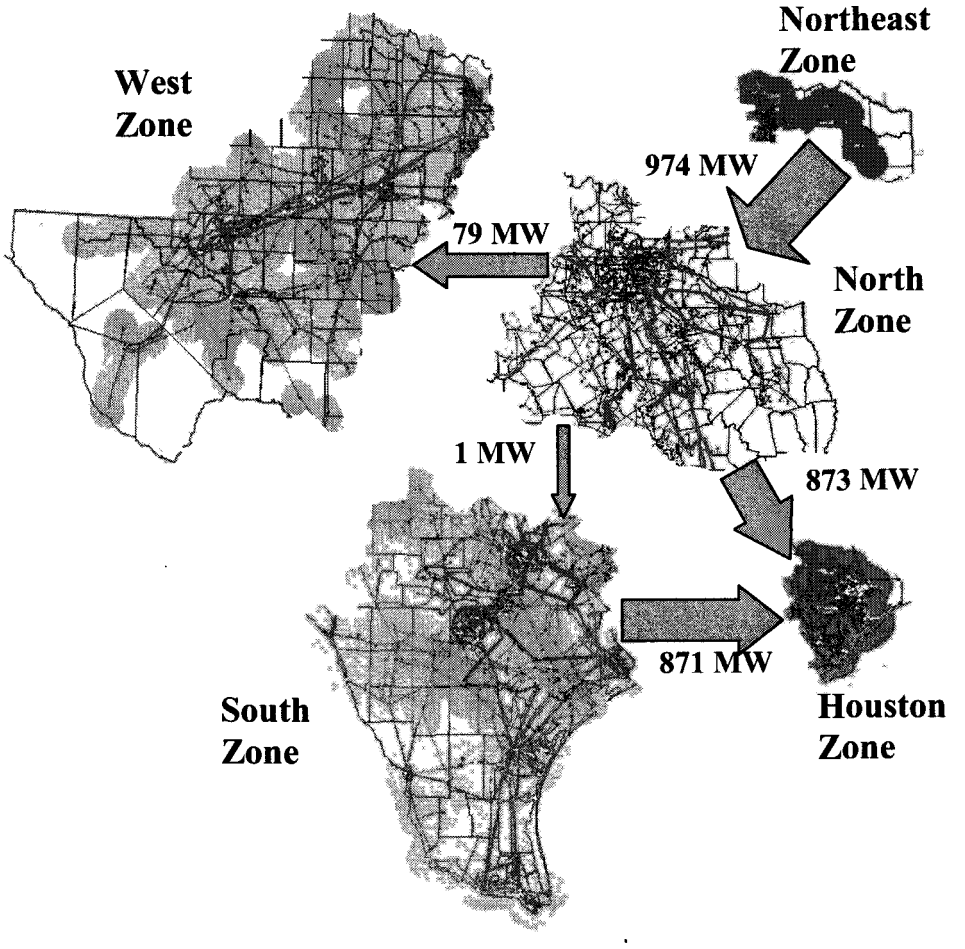
A. Electricity Flows between Zones

In 2006, there were five commercial pricing zones in ERCOT: (a) the North Zone, (b) the West Zone, (c) the South Zone, (d) the Houston Zone, and (e) the Northeast Zone, which was created in 2004 by dividing the North Zone. From year-to-year, slight adjustments are sometimes made to the boundaries of the commercial pricing zones, but the vast majority of customers remained in the same zone from 2005 to 2006. ERCOT operators use the SPD software to dispatch balancing energy in each zone in order to serve load and manage congestion between zones. The SPD model embodies the market rules and requirements documented in the ERCOT protocols.

To manage interzonal congestion, SPD uses a simplified network model with five zone-based locations and six transmission interfaces. These six transmission interfaces, referred to as Commercially Significant Constraints ("CSCs"), are simplified representations of groups of transmission elements. ERCOT operators use planning studies and real-time information to set limits for each CSC that are intended to utilize the total transfer capability of the CSC. In this subsection of the report, we describe the SPD model's simplified representations of flows between zones and analyze actual flows in 2006.

The SPD uses zonal approximations to represent complex interactions between generators, loads, and transmission elements. Because the model flows are based on zonal approximations, the estimated flows can depart significantly from real-time physical flows. Estimated flows that diverge significantly from actual flows are an indication of inaccurate congestion modeling leading to inefficient energy prices and other market costs. This subsection analyzes the impact of SPD transmission flows and constraints on market outcomes. Figure 52 shows the average SPD-modeled flows over CSCs between zones during 2006. A single arrow is shown for the modeled flows of both the North to West and West to North CSCs.

Figure 52: Average SPD-Modeled Flows on Commercially Significant Constraints During All Intervals in 2006



Note: In the figure above, CSC flows are averaged taking the direction into account. So one arrow shows the average flow for the for the North-to-West CSC was 79 MW, which is equivalent to saying that the average for the West-to-North CSC was *negative* 79 MW.

Figure 52 shows the five ERCOT geographic zones as well as the six CSCs that interconnect the zones: (a) the West to North interface, (b) the South to North interface, (c) the South to Houston interface, (d) the Northeast to North interface, (e) the North to Houston interface, and (f) the North to West interface. Based on SPD modeled flows, Houston is a significant importer while the Northeast Zone and the South Zone export significant amounts of power.

As discussed above, the simplified modeling assumptions specified in the ERCOT protocols for the current zonal market causes the interzonal power flows calculated by SPD to frequently diverge significantly from the actual flows. The most important simplifying assumption is that all generators in a zone have the same effect on the flows over the CSC, or the same generation shift factor ("GSF")[28] in relation to the CSC. In reality, the generators within each zone can have widely varying effects on the flows over a CSC. To illustrate this, we calculated flows that would occur over the CSC using actual generation and actual generation shift factors and compared this to flows calculated using actual generation and zonal average shift factors. Table 2 shows this analysis, which is based upon 2006 data but would not be significantly different for 2005. The flows over the North to West CSC are not shown separately in the table below since they are equal and opposite the flows for the West to North CSC.

Table 2: Average Calculated Flows on Commercially Significant Constraints Zonal-Average vs. Unit-Specific GSFs – 2006

CSC 2006	Flows Modeled by SPD (1)	Flows Calculated Using Actual Generation (2)	*Difference = (2) - (1)*	Flows Calculated Using Actual Generation and Unit-specific GSFs (3)	*Difference = (3) - (2)*
West-North	-79	-80	*-1*	-162	*-82*
South-North	-1	32	*33*	26	*-7*
South-Houston	871	874	*3*	1211	*337*
North-Houston	873	845	*-28*	661	*-184*
NorthEast-North	974	970	*-4*	942	*-28*

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

The first column in Table 2 shows the average flows over each CSC calculated by SPD. The second column shows the average flows over each CSC calculated using zonal-average GSFs and actual real-time generation in each zone instead of the scheduled energy and balancing energy deployments used as an input in SPD. Although these flows are both calculated using the same zonal-average GSFs, they can differ when the actual generation varies from the SPD generation. This difference is shown in the third column (in italics). These differences indicate that the actual generation levels result in higher calculated flows on each CSC except the West to North, North to Houston, and Northeast to North CSCs, where calculated flows are lower.

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

These results show that the heterogeneous effects of generators in a zone on the CSC flows can cause the actual flows to differ substantially from the SPD-calculated flows. Table 2 shows that the unit-specific GSFs increased the calculated flows on the South-Houston interface by 337 MW and reduced the calculated flows on the North to Houston CSC by 184 MW. These differences are sizable and are generally larger than the differences that can be attributed to variations in actual generation.

We note that the GSF simplification embedded in the SPD model is important for loads as well. Loads tend to be concentrated within a zone, but the SPD model assumes a generation-weighted average shift factor for all loads in the zone. Using generation-weighted shift factors for load rather than load-weighted shift factors can cause significant differences between SPD flows and actual flows. However, the impact of this assumption is diminished by the fact that loads are not used to manage transmission constraints in real-time. The use of simplified generation-weighted shift factors prevents the SPD model from efficiently assigning the costs of interzonal

congestion. In the long run, the use of generation-weighted shift factors for loads systematically biases prices, so that buyers in some zones pay too much, and others pay too little.

To effectively manage interzonal congestion, it is important for SPD to accurately model the major constrained transmission interfaces between zones. In 2006, the six CSCs modeled by SPD did not include all significant interfaces between zones. Sizeable quantities of power were transported on transmission facilities not modeled by SPD as flows on CSCs. Table 3 summarizes the actual net imports into each zone compared to SPD modeled flows from 2003 to 2006.

Table 3: Actual Net Imports vs. SPD-Calculated Flows on CSCs 2003 to 2006

Year	Zone	Actual Net Imports	SPD Flows on CSCs
2003	**Houston**	1796	565
	North	-507	191
	South	-1213	-702
	West	-76	-54
2004	**Houston**	2479	1265
	North	867	264
	NorthEast	-2116	-858
	South	-1531	-800
	West	304	129
2005	**Houston**	2596	1247
	North	660	164
	NorthEast	-2138	-845
	South	-1501	-728
	West	386	162
2006	**Houston**	3434	1744
	North	462	20
	NorthEast	-2334	-974
	South	-1741	-870
	West	180	79

Table 3 summarizes the differences between average SPD-calculated flows and average actual flows into each zone. These differences can be attributed to three factors. First, the use of zonal average GSFs, rather than resource-specific GSFs, by SPD to model generators can cause the SPD-calculated flows on a particular CSC to be substantially different from the actual flows.

Second, the use of generation-weighted shift factors to model load causes systematic differences between SPD flows and actual flows. For instance, SPD generally underestimated flows on the South to North CSC because of the difference between load-weighted and generation-weighted shift factors, accounting for a significant portion of the difference between SPD flows and net exports from the South Zone.

Third, significant quantities of power may flow over other transmission facilities that are not defined as part of the CSC. This will tend to cause the actual imports to exceed the SPD-calculated flows over the CSCs. For instance, the South-North interface is made up of the two 345 kV lines connecting the South and North zones, however, ERCOT has defined 19 CREs ("Closely Related Elements") which can also constrain flows from the South Zone to the North Zone. While ERCOT has the discretion to take CREs into account when managing interzonal congestion, they do not have the flexibility to do this efficiently. SPD always uses the CSC shift factors, although shift factors for CREs between the South Zone and North Zone may differ significantly from shift factors for the CSC. This leads to inefficient re-dispatch to manage constrained CREs.

Table 3 shows significant changes in the levels of net imports into each zone between 2003 and 2006. Imports to the Houston zone rose substantially from 2003 to 2004 and remained about the same from 2004 to 2005, followed by a steep increase again in 2006.[29] The West Zone shifted from being a net exporter in 2003 to importing substantial quantities in 2004 and in 2005, with the average import levels dropping by about 50 percent in 2006 compared to 2005. From 2003 to 2006, net exports increased from the South Zone as well as the combined area of the North and Northeast zones. In every case, the SPD-calculated flows on CSCs were significantly less than the actual interchange.

B. Interzonal Congestion

The prior subsection showed the average interzonal flows calculated by SPD compared to actual flows in all hours. This subsection focuses on those intervals when the interzonal constraints

[29] The North to Houston CSC was added in 2004.

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

Figure 53 shows the average SPD-calculated flows between the five ERCOT zones during constrained periods for the six CSCs. The arrows show the average magnitude and direction of the SPD-calculated flows during constrained intervals. The frequency with which these constraints arise is shown in parentheses.

Figure 53: Average SPD-Modeled Flows on Commercially Significant Constraints During Transmission Constrained Intervals in 2006

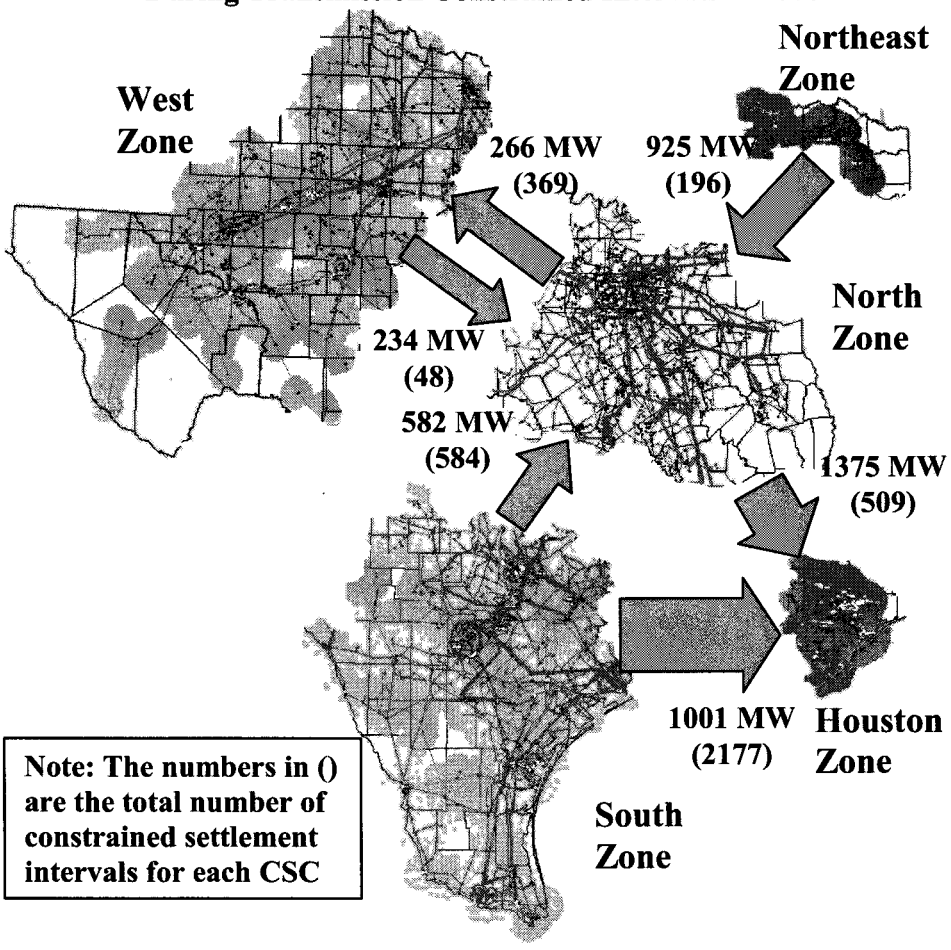


Figure 53 shows that inter-zonal congestion was most significant on the South to Houston CSC which exhibited SPD-calculated flows averaging 1,001 MW during 2,177 constrained intervals in 2006. Congestion was also significant on the South to North and North to Houston CSCs. The North to West CSC experienced much more congestion than the West to North CSC which

was congested for just 48 intervals during 2006. The Northeast to North CSC was constrained more frequently during 2006 than in 2005, although the majority of this congestion was related to transmission construction and the Northeast to North CSC was eliminated in 2007.

1. Congestion Rights in 2006

Interzonal congestion can be significant from an economic perspective, compelling the dispatch of higher-cost resources because power produced by lower-cost resources cannot be delivered over the constrained interfaces. When this occurs, participants must compete to use the available transfer capability between zones. To allocate this capability efficiently, ERCOT establishes clearing prices for energy in each zone that will vary in the presence of congestion and charges the transactions between the zones the interzonal congestion price.

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

To analyze the congestion rights in ERCOT, we first review the TCRs and PCRs that were allocated for each CSC in 2006. Figure 54 shows the average number of TCRs and PCRs that were allocated for each of the CSCs in 2006, as well as the average SPD-modeled flows during the constrained intervals.

Figure 54: Transmission Rights vs. Real-Time SPD-Calculated Flows Constrained Intervals – 2006

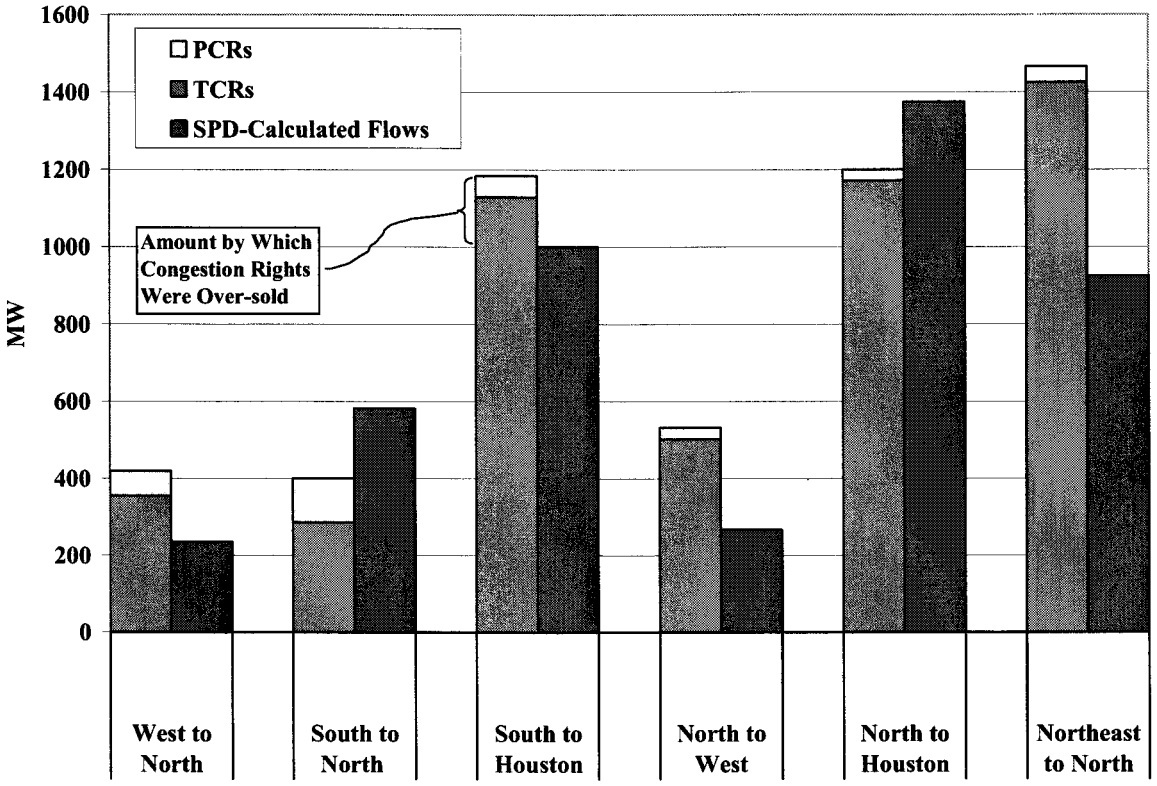


Figure 54 shows that total congestion rights (the sum of PCRs and TCRs) on the West to North, North to West, South to Houston, Northeast to North interfaces exceeded the average real-time SPD-calculated flows during constrained intervals while the congestion rights on the South to North and North to Houston CSCs were less than SPD calculated flows. These results indicate that the congestion rights were oversold in relation to the SPD-calculated limits for some CSCs. For instance, congestion rights for the South to Houston CSC were oversold by an average of 184 MW.

The largest divergence between the SPD-calculated limits and the limits implied by the congestion rights was on the Northeast to North CSC where 1,466 MW of congestion rights were allocated, but the average SPD-calculated flow during constrained intervals was 925 MW. Hence, the congestion rights that determine ERCOT's total obligation to make congestion payments exceeded the modeled flow over the CSC by an average of 541 MW.

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

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

To better understand the nature and causes of the shortfall implied by the results of Figure 54, we compare the SPD-calculated flows and congestion rights quantities for each of the constrained intervals by CSC.

2. Congestion on South to North CSC

Figure 55 shows the total quantity of congestion rights allocated by ERCOT for the South to North interface relative to the real-time SPD-calculated flows over the interface when the constraint was binding during 2006. Because only congested intervals are shown, some months will have significantly more observations than other months. Although some congestion occurred in every month, the three months from June to August accounted for 71 percent of all constrained intervals during 2006.

As explained in more detail below, the projected quantity of congestion rights changes from month to month as ERCOT reassesses the capability of each interface. ERCOT then adjusts the

quantity of TCRs accordingly in the monthly auctions. Figure 55 shows these changes in the congestion rights relative to the SPD-calculated flows, which fluctuate considerably in the congested intervals. In the figure, Total Congestion Rights include both TCRs and PCRs.

Figure 55: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to North – 2006

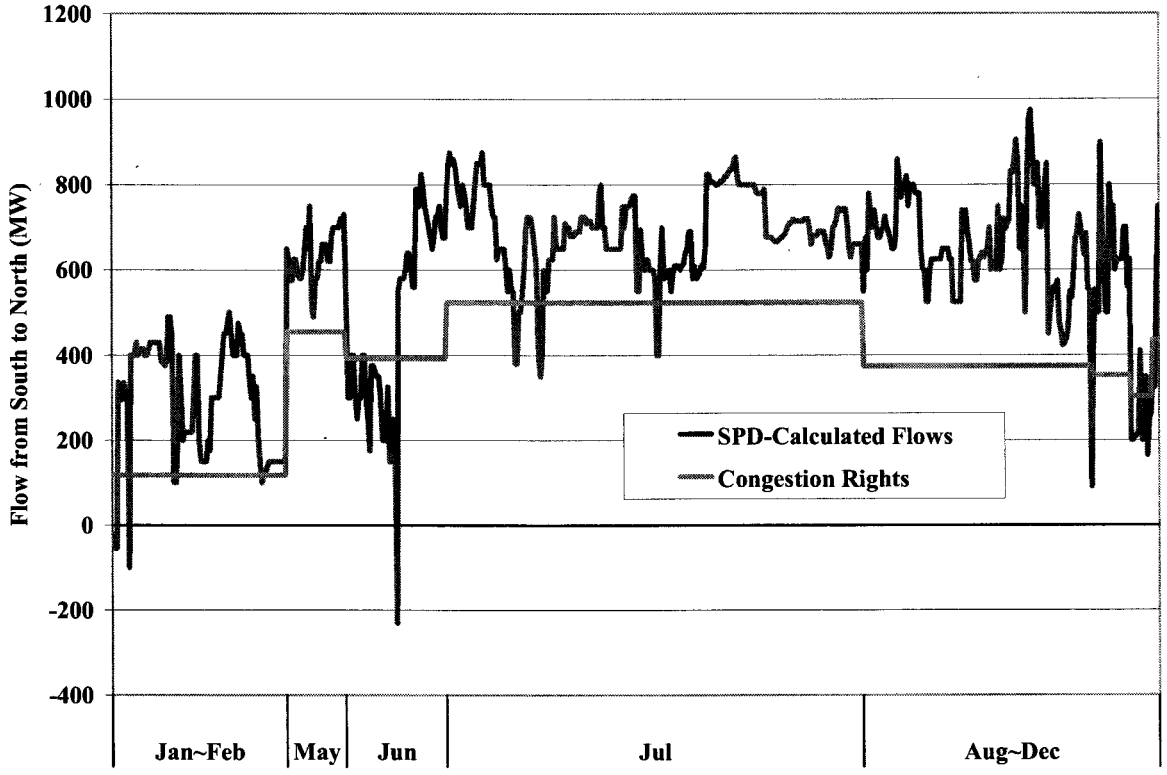


Figure 55 indicates that the quantity of outstanding congestion rights fluctuated considerably during 2006. From January to February, fewer than 200 MW of rights were allocated for the South to North CSC, whereas for May and July, more than 400 MW of congestion rights were allocated for the South to North CSC in 2006. This variation has to do with the complex nature of the South to North interface which results in it being constrained under a variety of circumstances.

Prior to each month, ERCOT estimates the transmission capability of the South to North interface based on transmission planning cases which use seasonal peak conditions. While two major lines make up the South to North interface, nearly 20 other transmission elements are defined as Closely Related Elements ("CREs"). Transmission constraints on the CREs can

reduce the amount that can be transferred across the two major lines. The pattern of flows can vary considerably, partly because of changes in the particular outages that are anticipated. Also, there is no guarantee that flows across the two main lines and all of the CREs will be in the same direction in every planning case. These issues highlight some of the problems that arise in the simplified zonal congestion management system. The nodal framework is better able to manage individual pieces of the transmission system, allowing more efficient utilization of the grid.

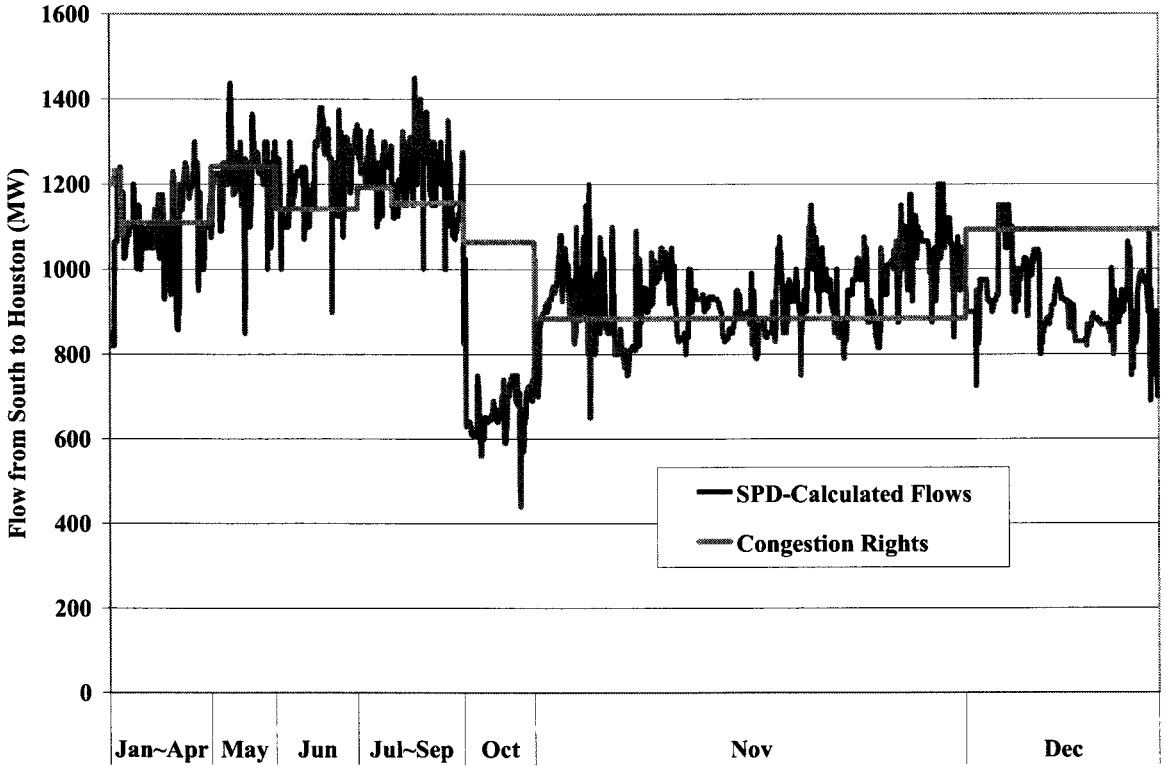
For the South to North CSC, the congestion rights were nearly always below SPD flows for the congested intervals in 2006. The figure shows five constrained intervals when the SPD-calculated flows were *negative* at times during January, February, and June.

These very low SPD-calculated flows generally do not reflect the actual physical flows in real time, *i.e.*, when the actual system conditions result in more flows over the South to North constraint than the simplified zonal model would predict. To prevent physical flows from exceeding the physical limits of the CSC, the ERCOT operators manually reduce the limit on the South to North interface in SPD. This causes SPD to redispatch generation in the various zones to reduce flows over the interface. Hence, because the SPD-calculated flows can be substantially different than actual flows, the ERCOT operators manage congestion by lowering the SPD limit when a constraint is physically binding to prevent additional flow over the CSC. Under extreme conditions, the operators must reduce the SPD limit into the negative range.

3. Congestion on South to Houston CSC

With 2,177 constrained intervals, this interface experienced the most frequent congestion of any CSCs during 2006. The most congestion occurred in November and December. In the months with significant congestion, SPD flows averaged between 940 and 924 MW. However, there was significant variation in the number of congestion rights allocated for this CSC by month, with as little as 884 MW in November and 1,241 MW in May. Figure 56 shows the comparison between actual flow and the congestion rights quantities.

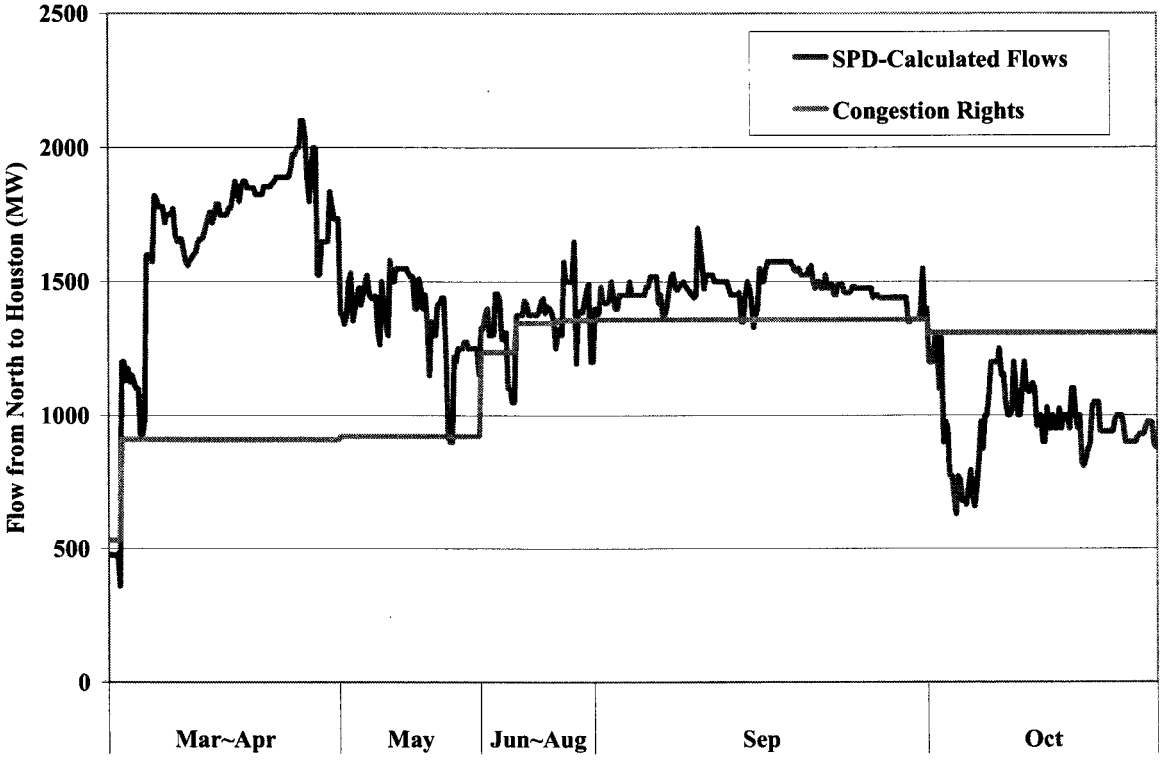
Figure 56: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to Houston – 2006



4. Congestion on North to Houston CSC

This CSC was created in 2004 to manage congestion on a path into Houston that is usually able to physically transfer more than 2,000 MW. Prior to May 2006, ERCOT generally allocated between 530 and 920 MW of congestion rights for this CSC. After May 2006, however, the number of congestion rights was increased to 1,235 in June and further to above 1,300 MW in subsequent months. From March to May, the rights were significantly under-sold while in October, the rights were significantly over-sold. Frequency of transmission constraints rose dramatically in September and October in conjunction with the increase of rights allocated.

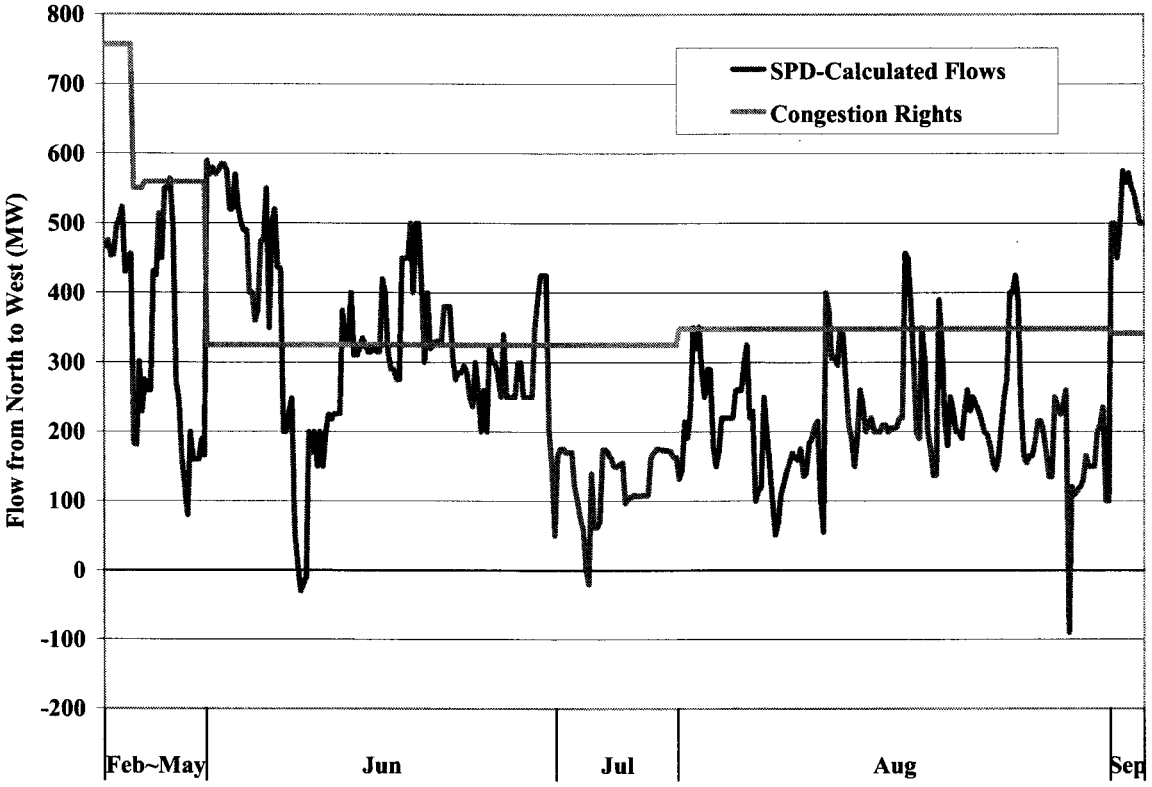
Figure 57: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals North to Houston – 2006



5. Congestion on North to West CSC

This CSC was congested primarily during the summer months with approximately 87 percent of constrained intervals in June, July and August. Although the number of congestion rights allocated for this interface varied from 325 to 757 MW over the year, the SPD flows averaged just 266 MW during constrained intervals.

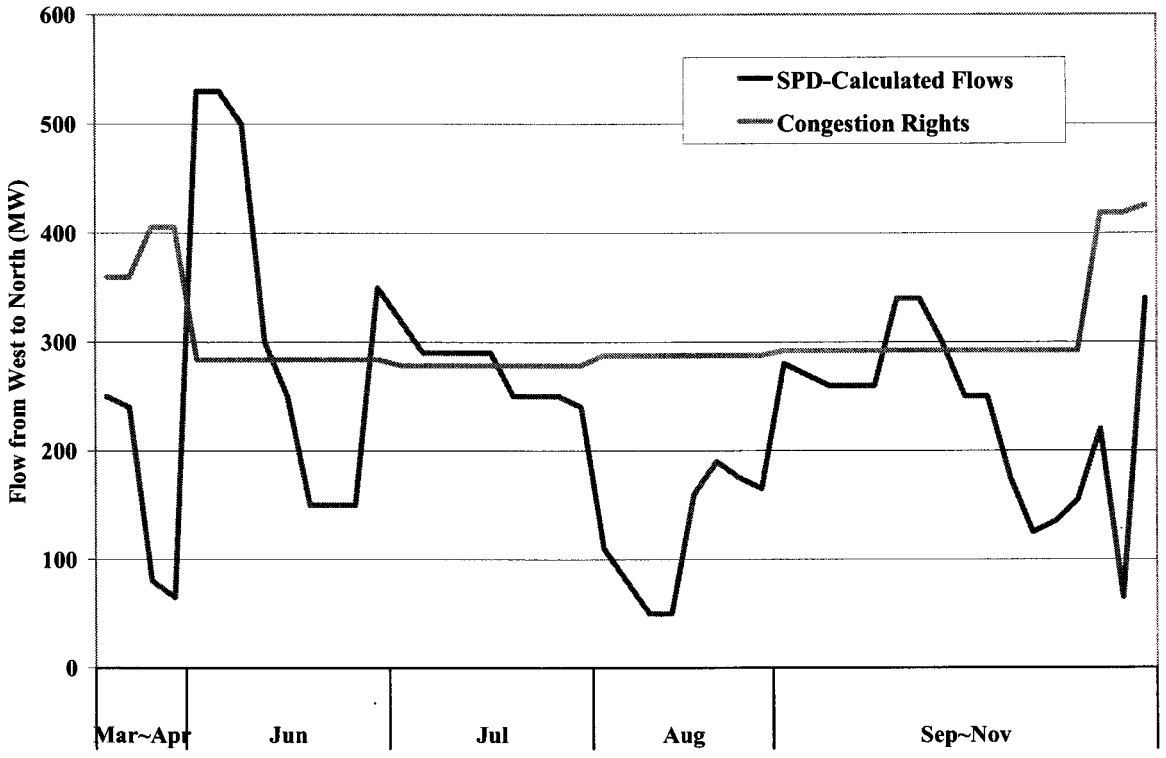
Figure 58: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals North to West – 2006



6. Congestion on West to North CSC

This CSC was the most infrequently congested CSC. During the year of 2006, the West to North CSC congested for only 48 intervals. This CSC was congested primarily during the summer months, June to September. Although the number of congestion rights allocated for this interface varied from 278 to 425 MW over the year, the SPD flows averaged just 234 MW during constrained intervals. As can be seen in Figure 59, in most of the months, the congestion rights were over sold.

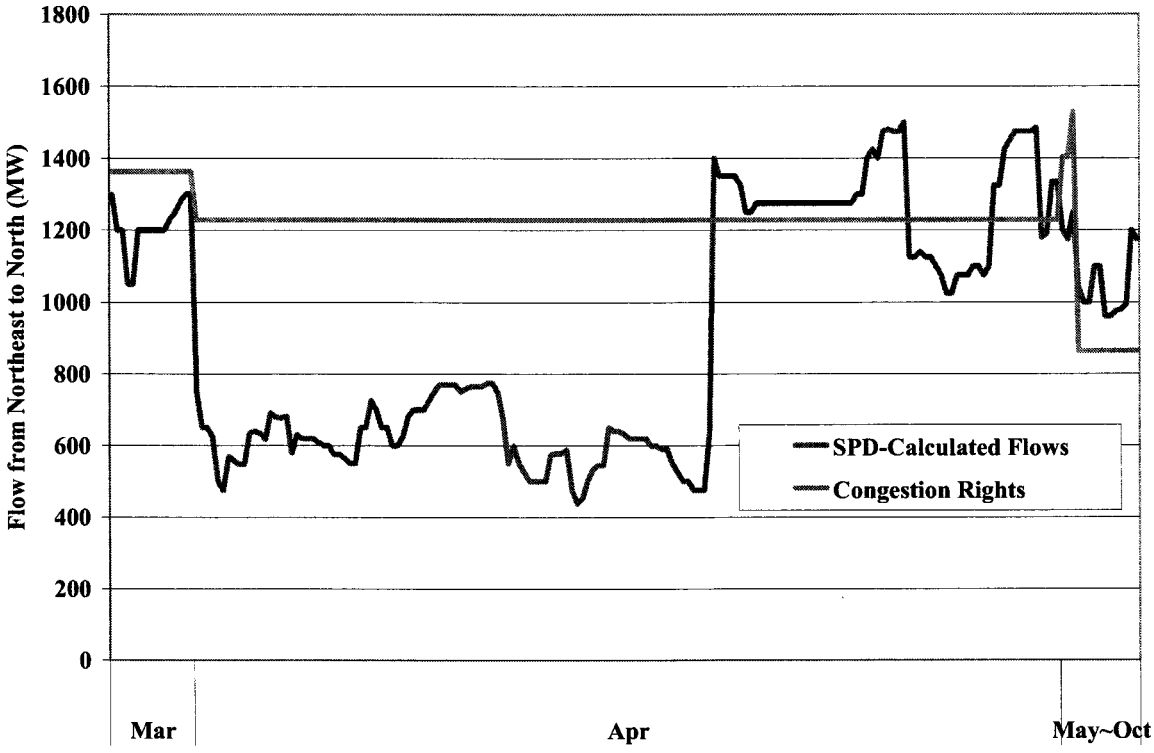
Figure 59: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals West to North – 2006



7. Congestion on Northeast to North CSC

The Northeast to North CSC was created in 2004. However, during the entire year of 2005, it was never congested. In 2006, the Northeast to North CSC congested 196 times with the actual flow between 438 to 1500 MW. During the months of March and April, the congestion rights exceed actual flow in most of the congested periods. Figure 65 shows the monthly comparison between actual flow and the number of congestion rights sold for the Northeast to North CSC. Most of the congestion on the Northeast to North CSC was associated with transmission construction related to increased transfer capability, and this CSC was eliminated in 2007.

Figure 60: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals Northeast to North – 2006



C. Congestion Rights Market

In this subsection, we review ERCOT's process to establish the quantity of congestion rights allocated or sold to participants. ERCOT performs transmission planning studies to determine the capability of each interface under peak summer conditions. This summer planning study is the basis for designating 40 percent of the transmission congestion rights sold in the annual auction. These rights are auctioned in December for the coming year. The remaining 60 percent of the transmission congestion rights are designated based on monthly updates of the summer study.[30] Since the monthly studies tend to more accurately reflect conditions that will prevail in

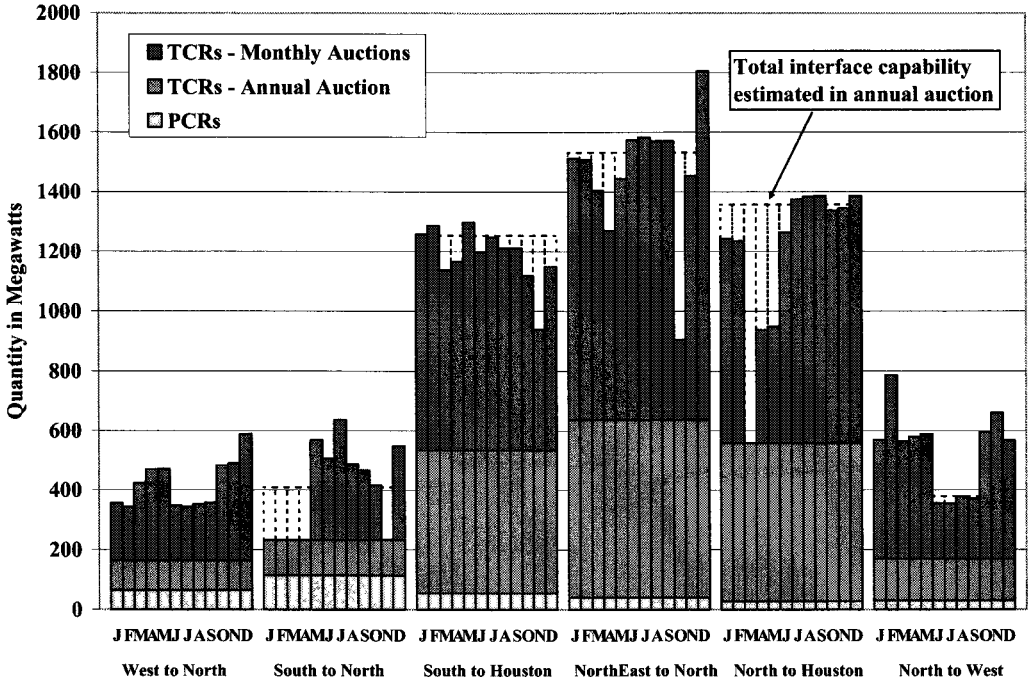
30 Prior to 2005, 60 percent of estimated capability (after accounting for Pre-assigned Congestion Rights which are assigned to NOIEs) was sold in the annual auction. The remaining 40 percent was sold in the monthly auctions. This was changed because there were instances when the capability estimated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

the coming month, the monthly designations tend to more closely reflect actual transmission limits.

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

To examine how these processes have together determined the total quantity of rights sold over each interface, Figure 61 shows the quantity of each category of congestion rights for each month during 2006. The quantities of PCRs and annual TCRs are constant across months and were determined before the beginning of 2006, while monthly TCR quantities can be adjusted monthly.

Figure 61: Quantity of Congestion Rights Sold by Type
2006



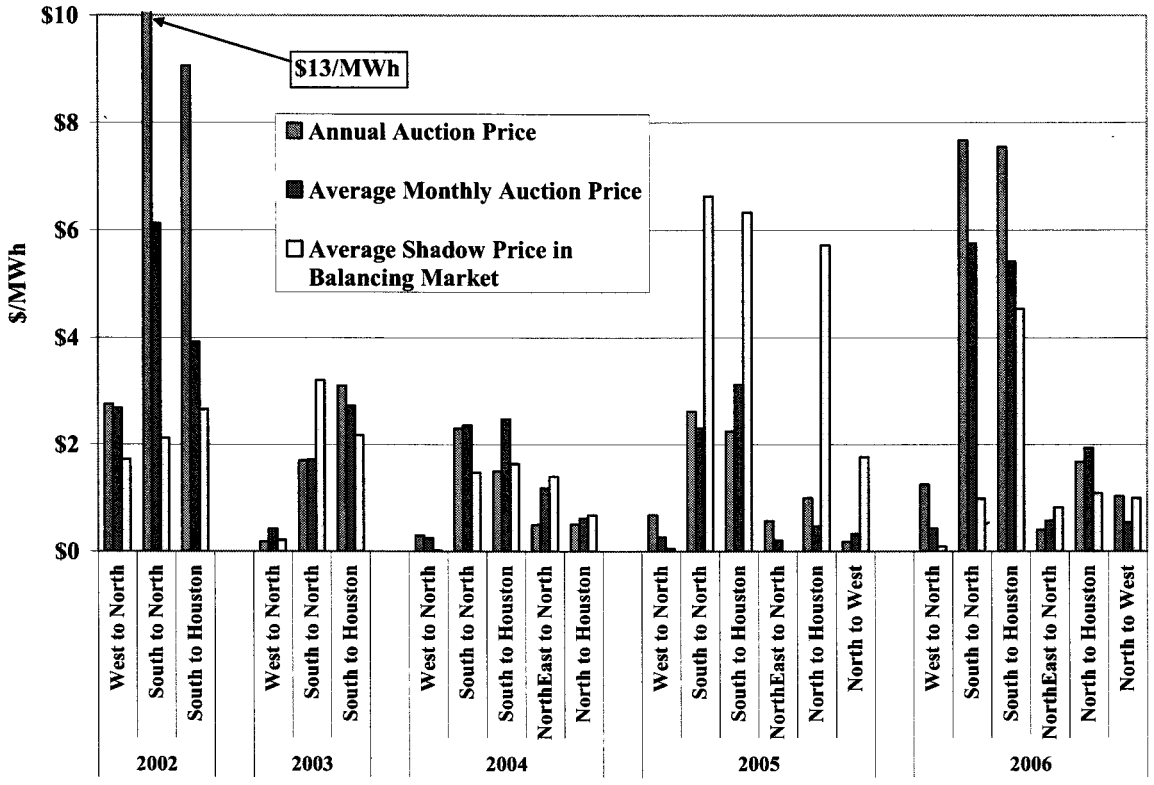
When the monthly planning studies indicate changes from the summer study, revisions are often made to the estimated transmission capability. Therefore, the auctioned congestion rights may increase or decrease relative to the amount estimated in the summer study. The shadow boxes in the figure represent the capability estimated in the summer study that is not ultimately sold in the monthly auction. When there is no shadow box in Figure 61, the total quantity of PCRs and TCRs sold in the annual and monthly auctions equaled or exceeded the summer estimate and therefore no excess capability is shown.

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

Market participants who are active in congestion rights auctions are subject to substantial uncertainty. Outages and other contingencies occur randomly that can substantially change the market value of a congestion right. Real-time congestion prices reflect the cost of interzonal congestion and are the basis for congestion payments to congestion rights holders. In a perfectly efficient system with perfect forecasting by participants, the average congestion price should equal the auction price. However, we would not expect full convergence in the real-world, given uncertainties and imperfect information. To evaluate the results of the ERCOT congestion rights market, in Figure 62 we compare the annual auction price for congestion rights, the average monthly auction price for congestion rights, and the average congestion price for each CSC.

Figure 62 indicates that in 2002, the annual auction for the TCRs resulted in prices that substantially over-valued the congestion rights, particularly on the South to North and South to Houston interfaces. Monthly TCR prices for these interfaces were roughly one-half of the prices from the annual auctions, but were still significantly higher than the ultimate congestion payments to the TCR holders. In the West to North interface, the annual and monthly TCR auction prices were close in magnitude and were both much closer to the true value of the congestion rights.

Figure 62: TCR Auction Prices versus Balancing Market Congestion Prices 2002 to 2006



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

In 2004, TCR auction prices for the West to North, South to North, and South to Houston interfaces were similar to the previous year. Since congestion tends to be consistent across time, the auction prices for 2004 were reasonable predictors of real-time congestion. In 2004, there were two new products in the TCR auctions for the new Northeast to North and North to Houston CSCs. In both cases, the annual TCR price was below the monthly average TCR price, which was slightly below the average value of congestion, but the divergence between auction prices and actual congestion values was not as significant as in 2002. This reflects cautiousness

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

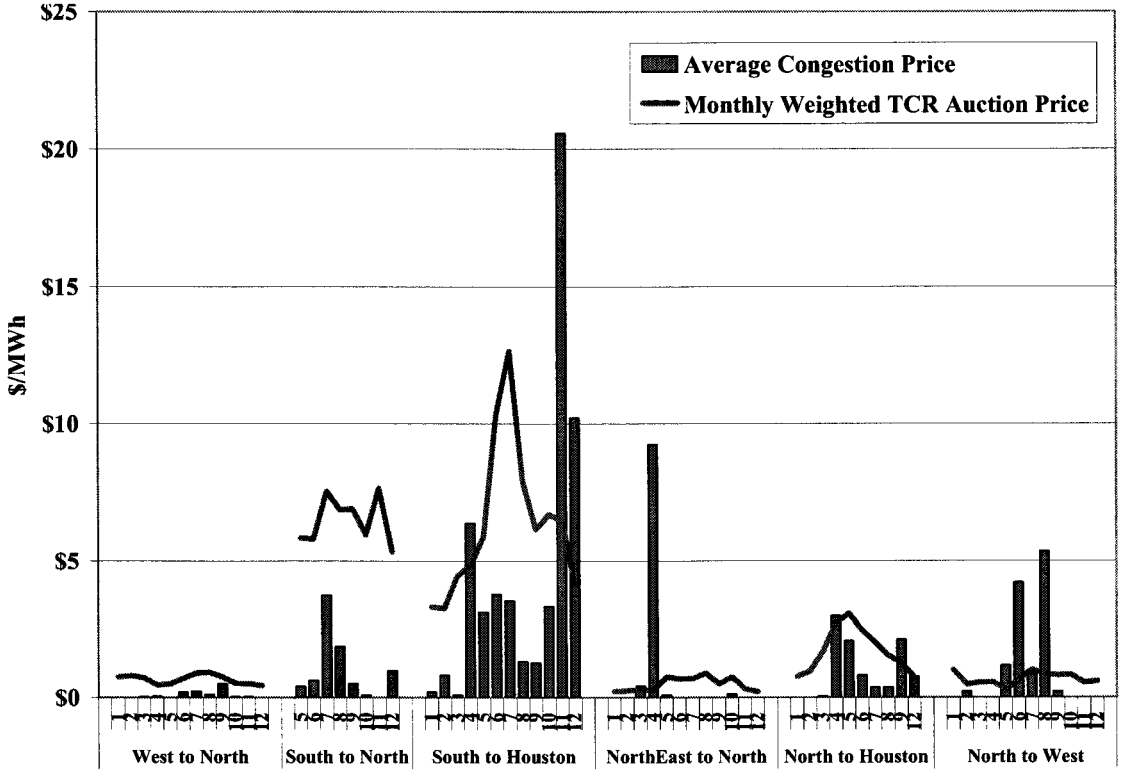
In 2005, market participants substantially under-estimated the value of congestion on the CSCs. The annual and monthly TCR prices in 2005 were generally in line with the TCR prices and the levels of balancing market shadow prices that prevailed in 2004. However, the actual volume and prices of congestion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

In contrast to 2005, market participants over-estimated the annual value of congestion on the South to North, South to Houston, and North to Houston CSCs in 2006. The annual auction price for the North to West CSC converged well with the congestion value. The West to North CSC congestion value was again over-estimated as was the case from 2002 to 2005. Although South to North TCR auction concluded with the highest auction price among all the other CSCs, the South to North CSC actual congestion value decreased significantly from 2005.

Figure 63 compares monthly TCR auction prices with monthly average real-time CSC shadow prices from SPD for 2006. The TCR auction prices are expressed in dollars per MWh. In months when the monthly auction did not occur (*i.e.*, when the annual auction designated sufficient congestion rights for that month) no data is presented. This explains the missing months for the South-North CSC and the North-Houston CSC.[31]

[31] Notice that these missing months correspond to the missing monthly auction values in Figure 61.

Figure 63: Monthly TCR Auction Price and Average Congestion Value 2006



The TCR price trends for South to North and North to Houston CSCs, correlated well with the actual congestion prices, although the TCR prices for the South to North CSC far exceeded the congestion prices. Overall, market participants did a poor job predicting fluctuations in congestion during 2006, particularly on the South to Houston and Northeast to North interfaces. For both of these interfaces, there were several months when balancing market congestion spiked, far exceeding the TCR prices in those months. However, based on the TCR prices, there is little sign that market participants expected an increase in congestion in those months relative to other months.

To evaluate the total revenue implications of the issues described above, our next analysis compares the TCR auction revenues and obligations. Auction revenues are paid to loads on a load-ratio share basis. Market participants acquire TCRs in the ERCOT-run TCR auction market in exchange for the right to receive TCR credit payments (equal to the congestion price for a CSC times the amount of the TCR). If TCR holders could perfectly forecast shadow prices in the balancing energy market, auction revenues would equal credit payments to TCR holders.

The credit payments to the TCR holders should be funded primarily from congestion rent collected in the real-time market from participants scheduling transfers between zones or power flows resulting from the balancing energy market.

The congestion rent from the balancing energy market is associated with the schedules and balancing deployments that result in interzonal transfers during constrained intervals (when there are price differences between the zones). For instance, suppose the balancing energy market deployments result in exports of 600 MWh from the West Zone to the North Zone when the price in the West Zone is $40/MWh and the price in the North Zone is $55/MWh. The customers in the North Zone will pay $33,000 (600 MWh * $55/MWh) while suppliers in the West Zone will receive $24,000 (600 MWh * $40/MWh). The net result is that ERCOT collects $9,000 in congestion rent ($33,000 – $24,000) and uses it to fund payments to holders of TCRs.[32] If the quantity of TCRs perfectly matches the capability of the CSC in the balancing energy market, the congestion rent will perfectly equal the amount paid to the holders of TCRs.

Figure 64 reviews the results of these processes by showing (a) monthly and annual revenues from the TCR auctions, (b) credit payments earned by the holders of TCRs based on real-time outcomes, and (c) congestion rent from schedules and deployments in the balancing energy market.

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

Figure 64: TCR Auction Revenues, Credit Payments, and Congestion Rent[33] 2002 to 2006

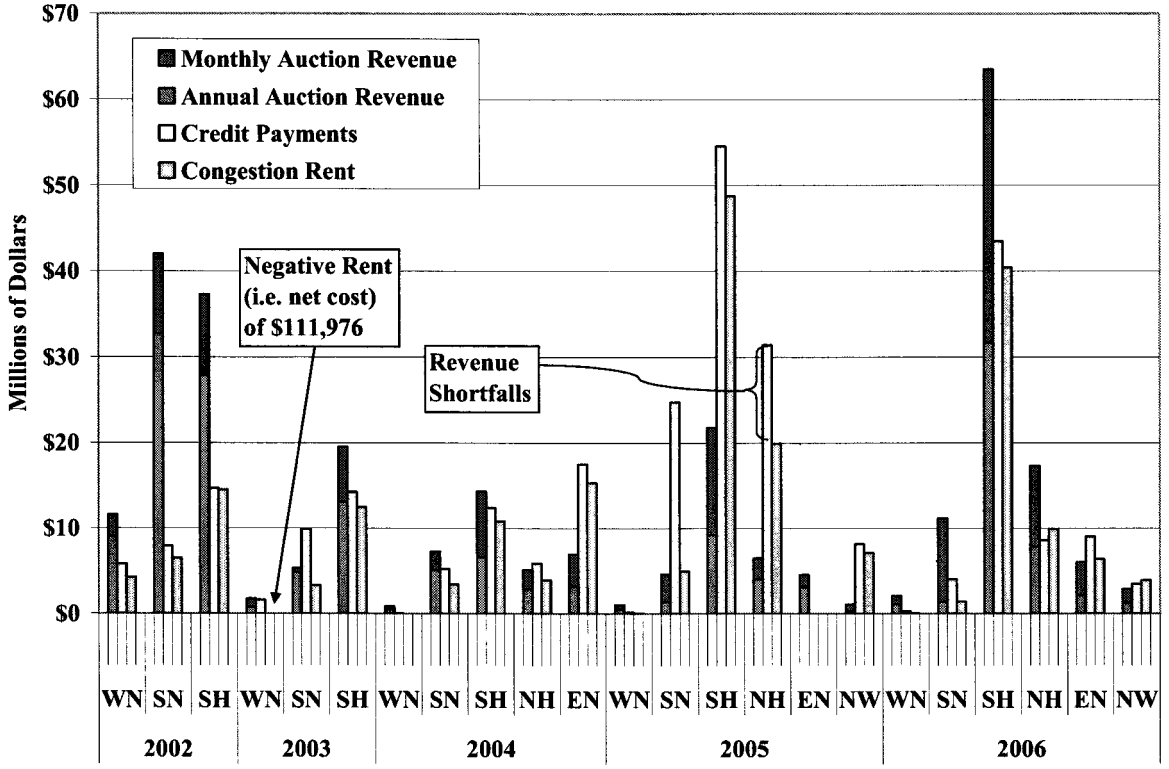


Figure 64 shows that in 2002, the total auction revenues were far greater than credit payments to TCR holders. This is the result of the auction prices being much greater than the average shadow prices that occurred in the balancing energy market (as was shown in Figure 63 above). The figure also shows that from 2002 to 2003, there was a significant reduction in auction revenues (a reduction of 71 percent). Auction revenues were reduced in 2003 because both annual and monthly auction prices decreased significantly due to improvements in the ability of market participants to forecast congestion on CSCs.

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

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

substantially under-valued congestion on the Northeast to North interface, which was also new in 2004.

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

In contrast to 2005, auction revenues for the South to North, South to Houston and North to Houston interfaces exceeded credit payments in 2006. As shown in Figure 63, for those interfaces, auction prices exceeded the congestion prices. The magnitude of credit payments are in the same trend as in 2005, but the 2006 South to North and North to Houston interfaces exhibited far less credit payments and congestions rent compared to 2005. Northeast to North interfaces experienced more congestion than 2005 and hence the credit payments went up compared to 2005.

Figure 64 also shows that payments to TCR holders have consistently exceeded the congestion rents that have been collected from the balancing market since the creation of the TCR market. The difference was relatively modest in 2002 when congestion rents covered 93 percent of payments to TCR holders and in 2004 when they covered 81 percent. However, in 2003 and 2005, congestion rents covered only 61 percent and 68 percent, respectively, of payments to TCR holders. In 2006, congestion rents covered 90 percent of payments to TCR holders, which is an improvement from previous years. When congestion rents fall significantly below payments to TCR holders, it implies that the SPD-calculated flows across constrained interfaces have been systematically lower than the amount of TCRs sold for the interfaces.

As described above, a revenue shortfall exists when the credit payments to congestion rights holders exceed the congestion rent. This shortfall is caused when the quantity of congestion

rights exceeds the SPD-calculated flow limits in real-time.[34] These shortfalls are included in the Balancing Energy Neutrality Adjustment charge and assessed to load ERCOT-wide. Collecting substantial portions of the congestion costs for the market through such uplift charges reduces the transparency and efficiency of the market. It also increases the risks of transacting and serving load in ERCOT because uplift costs cannot be hedged.

D. Local Congestion and Local Capacity Requirements

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

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

---

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

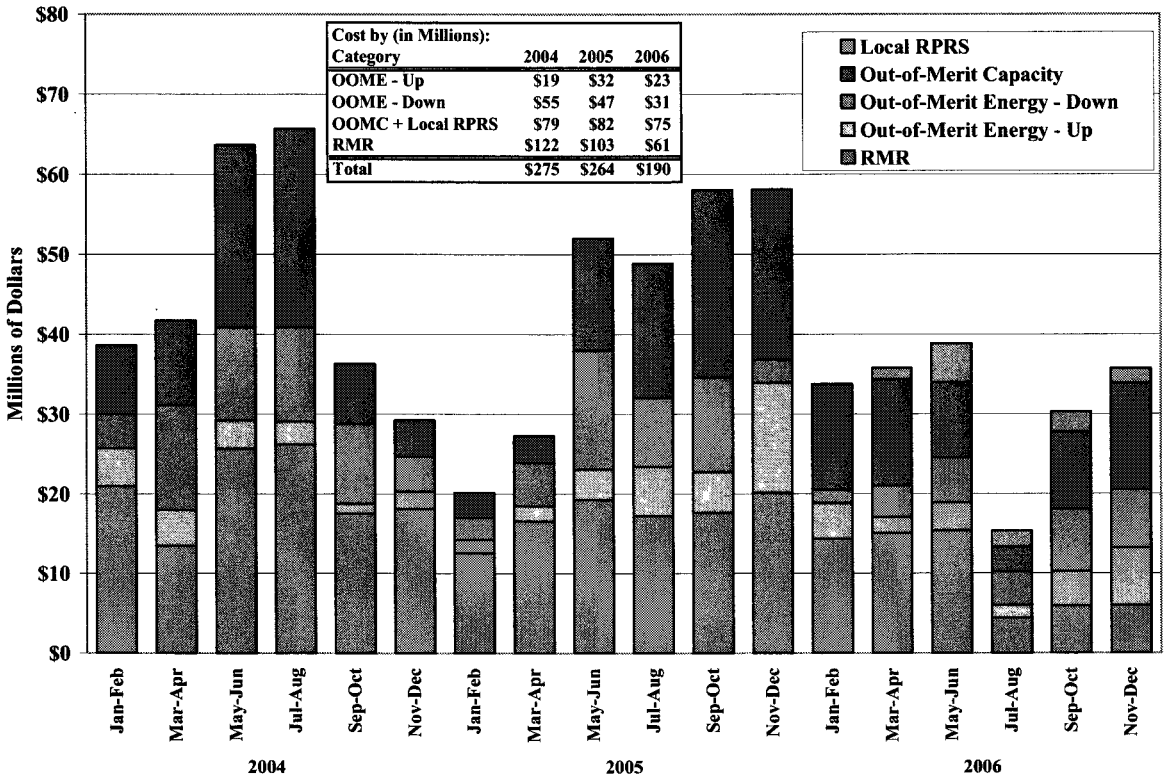
When a unit is dispatched out of merit (OOME up or OOME down), the unit is paid for a quantity equal to the difference between the scheduled output based on the unit's resource plan and the actual output resulting from the OOME instruction from ERCOT. The payment per MWh for OOME is a pre-determined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

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

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

Finally, RMR units committed or dispatched pursuant to their RMR agreements receive cost-based compensation. Since October 2002, ERCOT has entered into several RMR agreements with older, inefficient units that were planned to be retired. However, as a part of the RMR exit strategy process, all but three units were removed from RMR status by mid-2006. Units contracted to provide RMR service to ERCOT are compensated for start-up costs, energy costs, and are also paid a standby fee. Figure 65 shows each of the four categories of uplift costs from 2004 to 2006.

Figure 65: Expenses for Out-of-Merit Capacity and Energy 2004 to 2006



The results in Figure 65 show that overall uplift costs for RMR units, OOME units, and OOMC/Local RPRS units were relatively consistent between 2004 and 2005. The costs decreased by $74 million in 2006 from $264 million to $190 million, a reduction of 28 percent. As previously noted, there were substantial reductions to RMR cost due to the expiration of RMR agreements in 2006, which accounts for $42 million of the $74 million decrease from 2005 to 2006. Total OOME Up and OOME Down costs also decreased from $79 million in 2005 to $54 million in 2006, a reduction of 32 percent. This reduction is likely due to the continued improvements to the ERCOT transmission system resulting in less frequent local congestion, and the introduction of an enhanced replacement reserve procurement process by ERCOT in 2006.

Although the costs are borne by load throughout ERCOT, the costs are caused in specific locations because these actions, with the exception of zonal RPRS, are taken to maintain local reliability. The rest of the analyses in this section evaluate in more detail where these costs were caused and how they have changed between 2004 and 2006. Figure 66 shows these payments by location.