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 36 and Figure 37). This analysis is similar to that shown in Figure 17 and Figure 18, except instead of showing balancing energy prices relative to load, we show balancing energy prices relative to balancing energy deployments. Figure 38 shows the analysis for the ramping-up hours.

Figure 38: Balancing Energy Prices and Volumes
Ramping-Up Hours

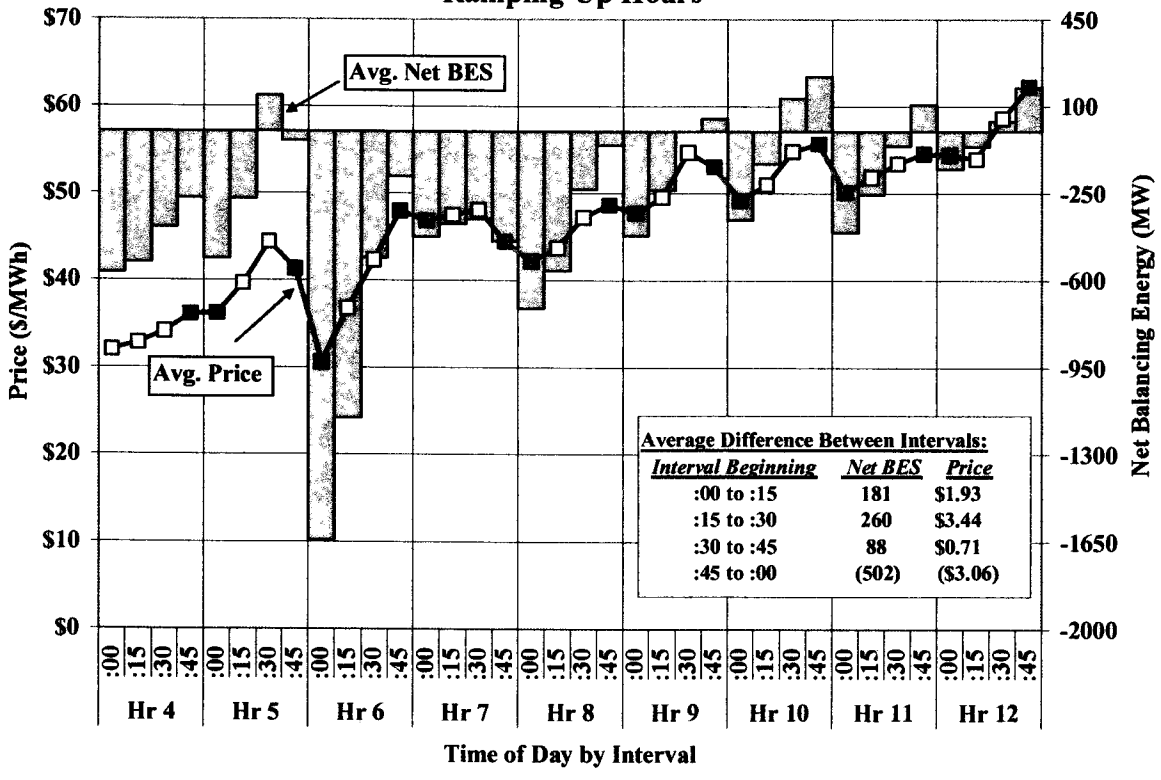
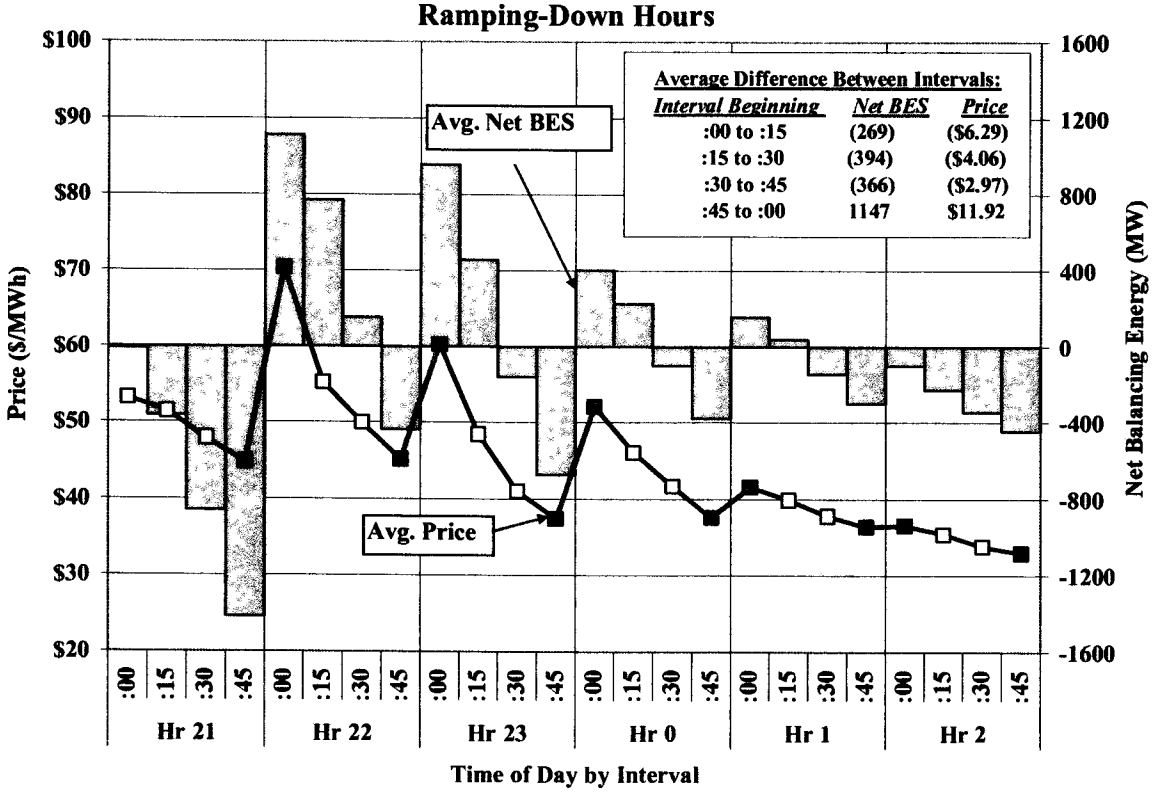


Figure 38 reveals two key aspects of the balancing energy market. First, as discussed above, balancing energy prices are highly correlated with balancing energy deployments. Second, with

the exception of hour 7 and 9, 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 39 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 will largely resolve these inefficiencies.

Figure 39: Balancing Energy Prices and Volumes Ramping-Down Hours



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 38 and Figure 39.

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 2007. 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.

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.[21] The operational implications associated with these issues continued in 2007 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.

C. Balancing Energy Market Offer Patterns

In this section, we evaluate balancing energy offer patterns by analyzing the rate at which capacity is offered.[22] Figure 40 shows the average amount of capacity offered to supply balancing up service relative to all available capacity.

21 2005 SOM Report at 68-76.

22 The methodology for determining the quantities of un-offered capacity is detailed in the 2006 SOM Report (2006 SOM Report at 63-65).

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

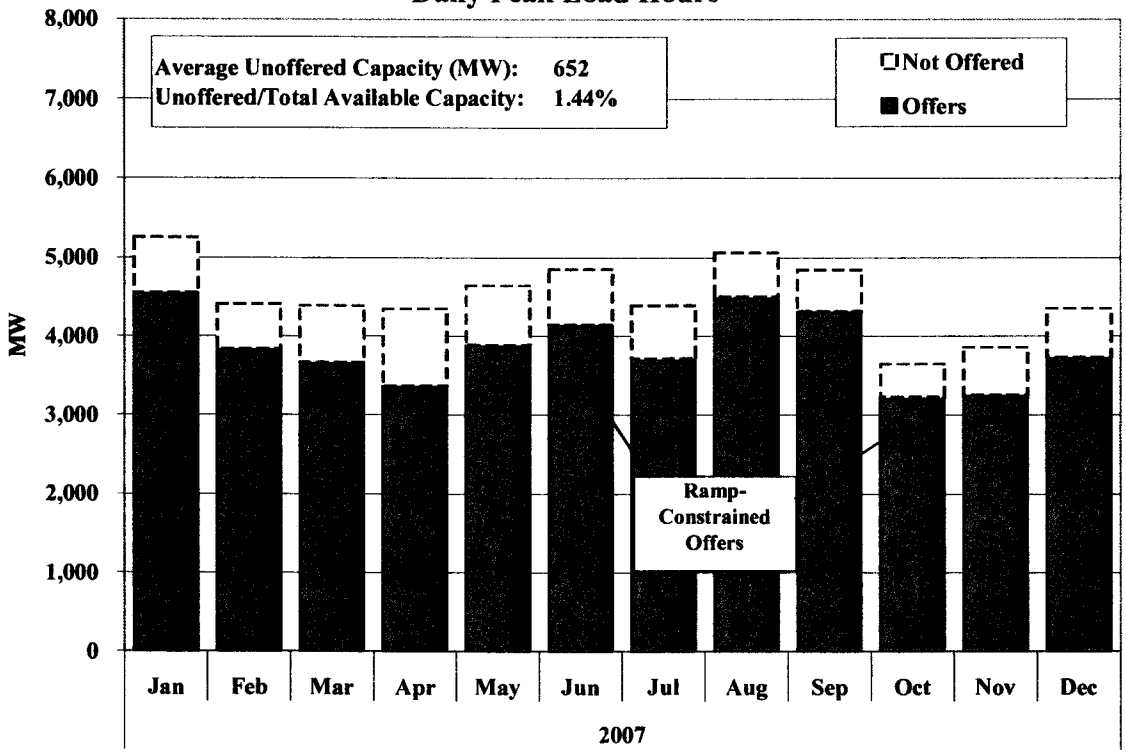
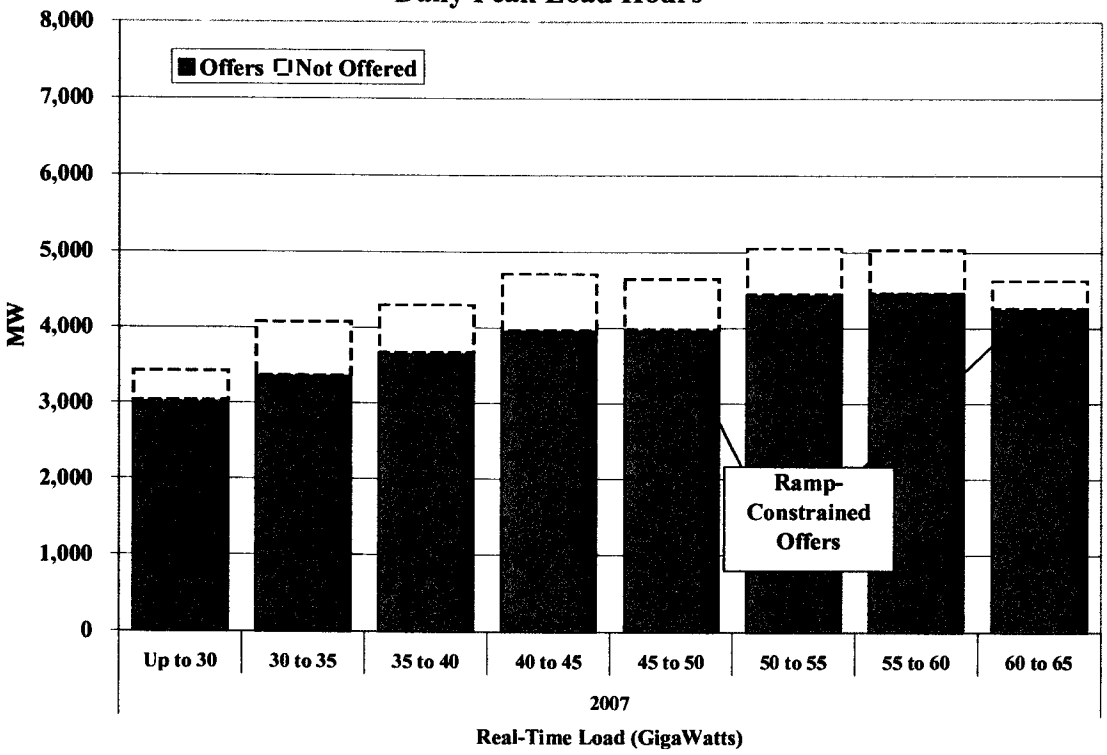


Figure 40 shows only slight variation in 2007 over time in quantities of energy available and offered to the balancing energy market. Up balancing offers are divided into the portion that is capable of being deployed in one interval and the portion which would take longer due to portfolio ramp rate offered 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 41 shows the same data as the previous figure, but arranged by load level for daily peak hours in 2007. 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 41: Balancing Energy Offers Compared to Total Available Capacity Daily Peak Load Hours



The figure indicates that in 2007 the average amount of capacity available to the balancing market increased gradually up to 60 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 41 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.[23]

[23] *See* 2006 SOM Report at 67 for a discussion of the residual un-offered capacity.

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 2007 and the existing generating capacity available to satisfy the load and operating reserve requirements.

A. ERCOT Loads in 2007

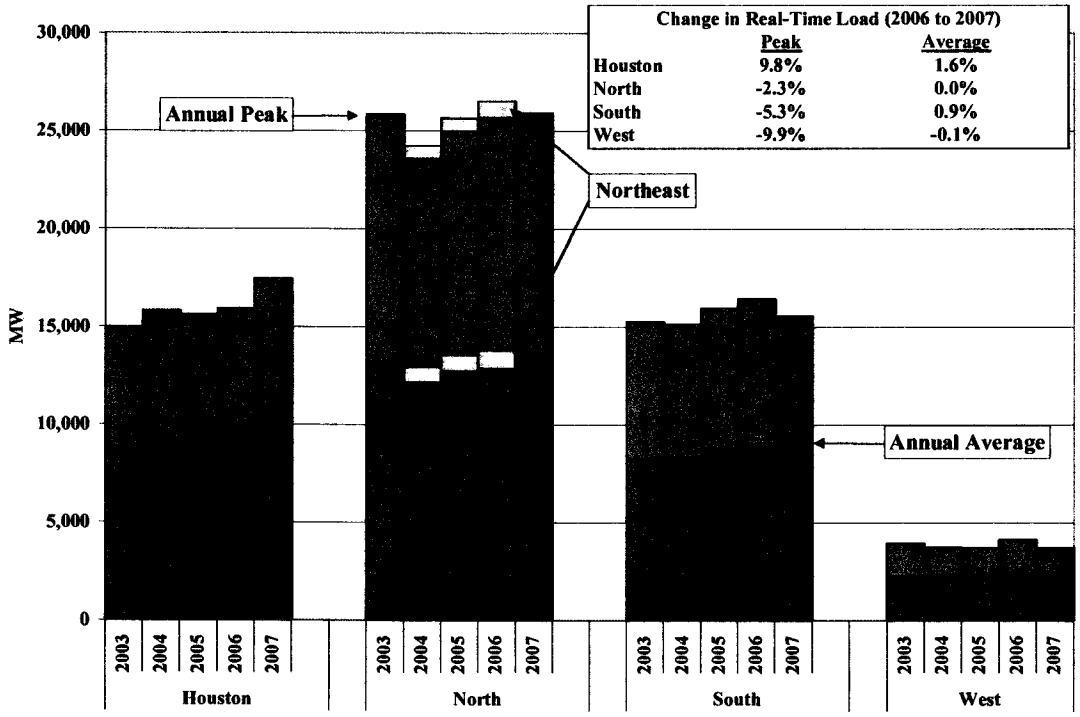
There are two important dimensions of load that should be evaluated separately. First, the changes in overall load levels from year to year can be shown by tracking the changes in average load levels. This metric will tend to capture changes in load over a large portion of the hours during the year. 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 2007 are examined in this subsection and summarized in Figure 42.[24]

This figure shows peak load and average load in each of the ERCOT zones from 2003 to 2007. 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 40 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 is the smallest (with about 7 percent of the total ERCOT load). Figure 42 shows the annual non-coincident peak load for each zone. This is the highest

[24] The load values in this Section are from ERCOT settlement data. In previous State of the Market Reports, the load values were from ERCOT's Scheduling, Pricing and Dispatch software (including transmission and distribution losses). Data from 2003 to 2006 have also been adjusted.

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 zones was greater than the annual ERCOT peak load.

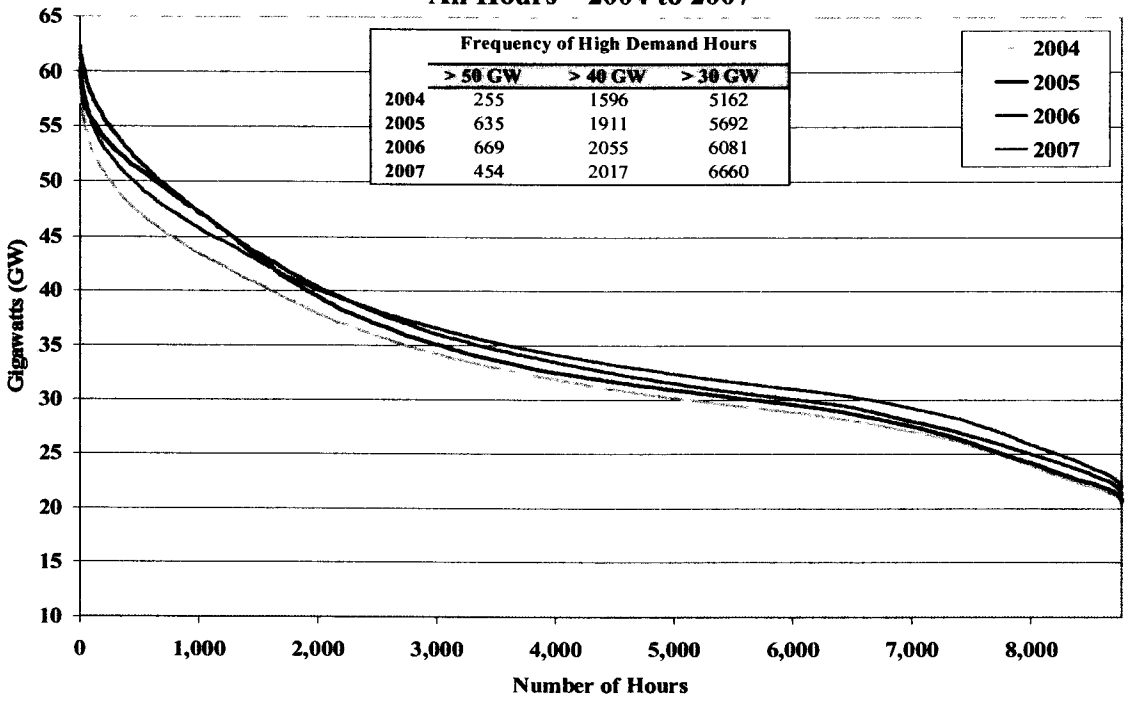
Figure 42: Annual Load Statistics by Zone
2003 to 2007



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 43 compares load duration curves for each year from 2003 to 2007. A load duration curve shows the number of hours (shown on the horizontal axis) that load exceeds a particular level (shown on the vertical 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.

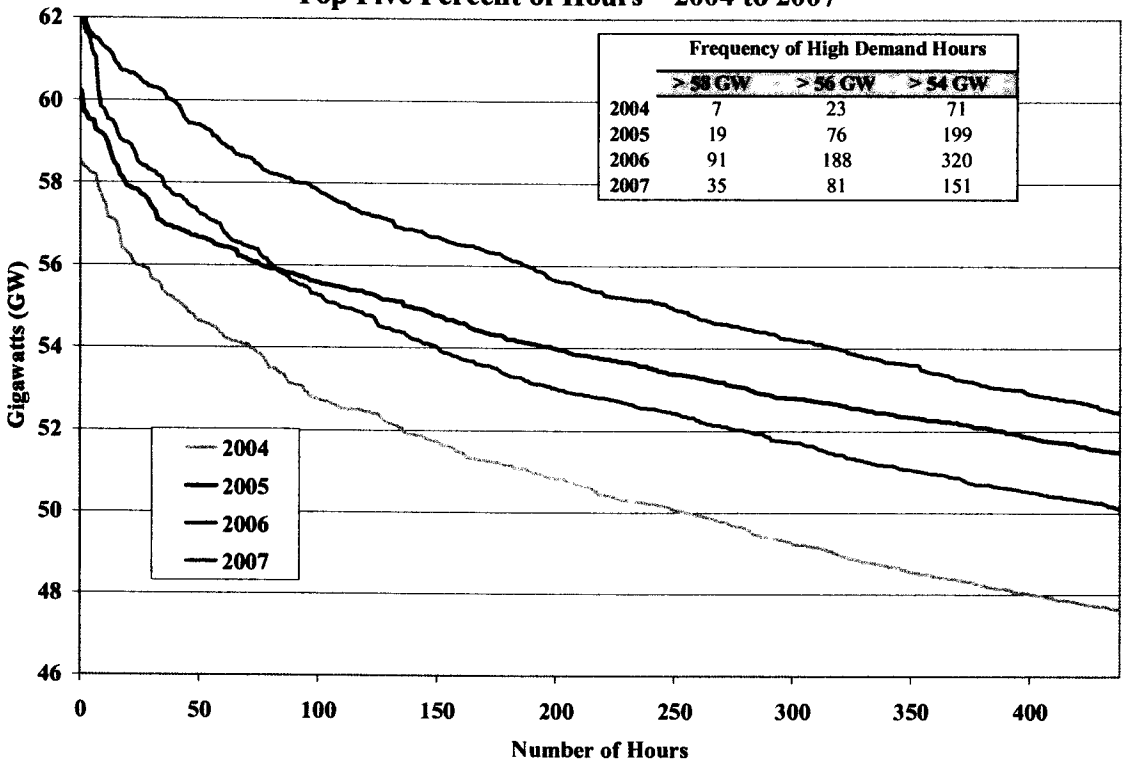
Figure 43: ERCOT Load Duration Curve
All Hours – 2004 to 2007



As shown in Figure 43 , the load duration curve for 2007 lies above the curves for the previous four years at load levels less than 40 GW. Load increased about 0.7 percent from 2006 to 2007. In 2007, there were 10 percent more hours when load exceeded 30 GW than in 2006.

To better show the differences in the highest-demand periods between years, Figure 44 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 2003 to 2006, the frequency of high demand hours in 2007 dropped compared with year 2006. Load exceeded 58 GW in 35 hours in 2007, 91 hours in 2006, 19 hours in 2005, 7 hours in 2003 and 8 hours in 2004. The same pattern prevailed at lower load levels.

Figure 44: ERCOT Load Duration Curve
Top Five Percent of Hours – 2004 to 2007

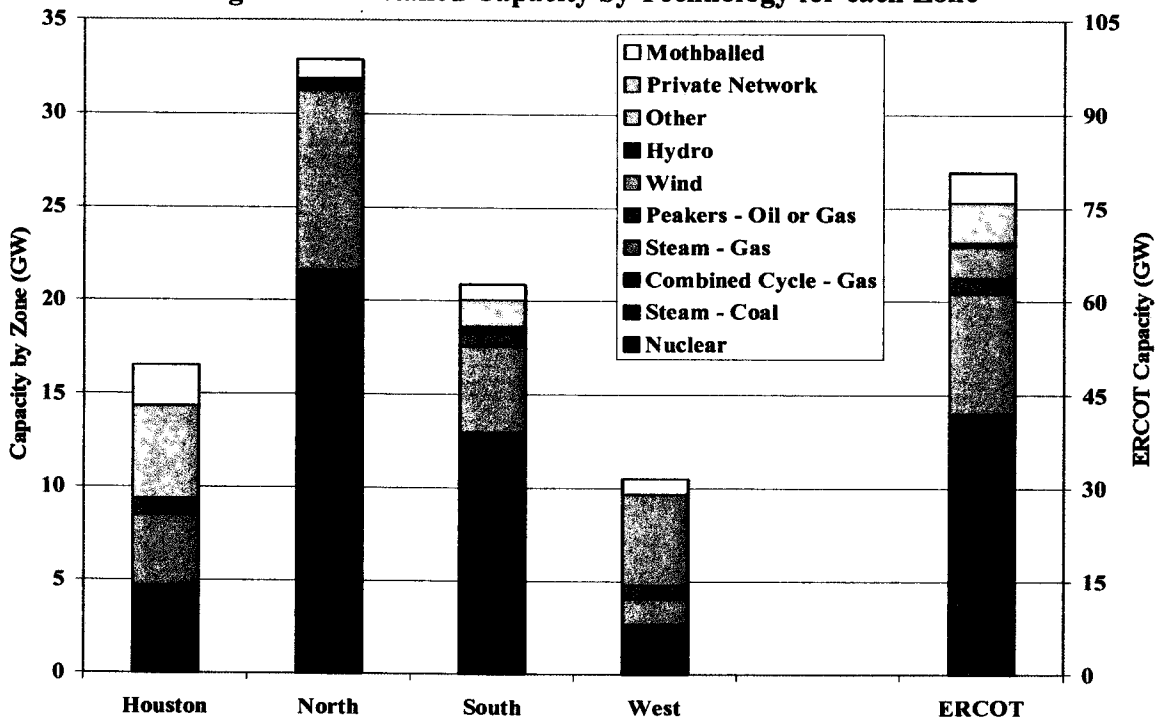


This figure also shows that the peak load in each year was roughly 15 to 25 percent greater than the load at the 95th percentile of hourly load. For instance, in 2006, the peak load value was over 62 GW while the 95th percentile was about 52 GW. This is typical of, and even somewhat flatter than, the load patterns in most electricity markets. This implies that a substantial amount of capacity, more than 10 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 45 shows the installed generating capacity by type in each of the ERCOT zones.

Figure 45: Installed Capacity by Technology for each Zone



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 70 percent of generation capacity in ERCOT as a whole, and 85 percent in the Houston Zone. Much of this natural gas-fired capacity represents relatively new combined-cycle units than have been installed throughout ERCOT over the past 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.

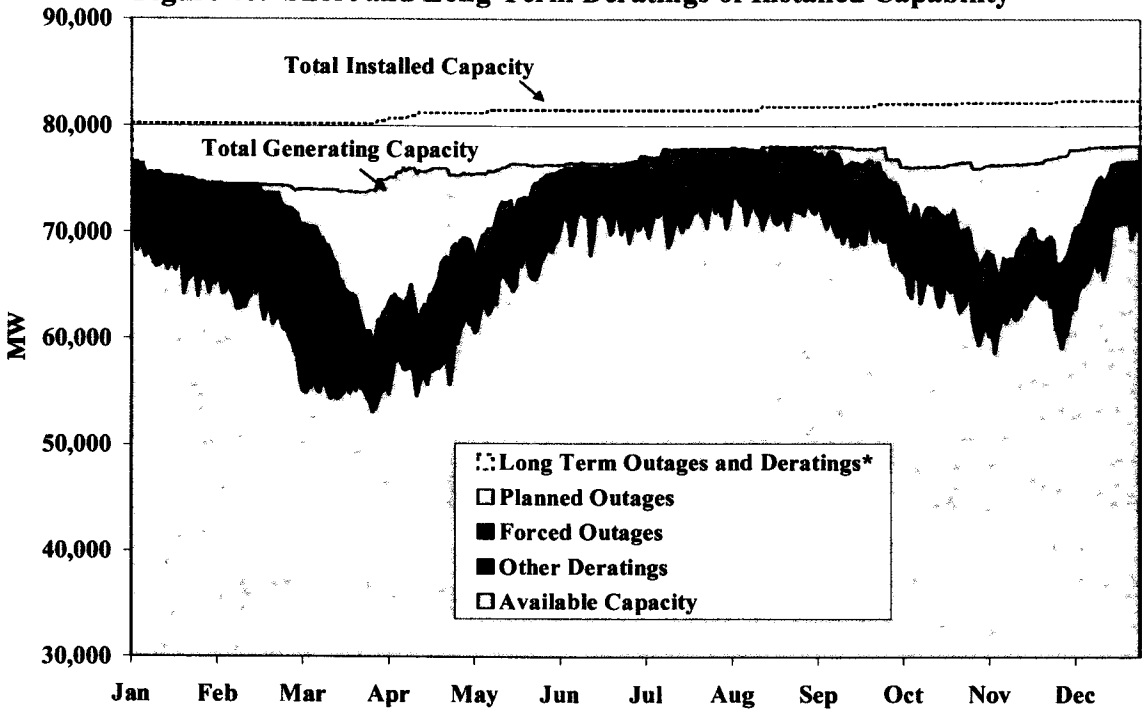
The distribution of capacity among the ERCOT zones is similar to the distribution of demand. This is consistent with the legacy of investment under the regulated vertically integrated utilities when load and resources were largely integrated within separate control areas. The North Zone accounts for 38 percent of capacity, the South Zone 28 percent, the Houston Zone 22 percent, and the West Zone 11 percent. The Houston is typically an importer of power, while the North and South Zones typically export power. Because large amounts of power flow out of the South and the North Zones into the Houston Zone, the South-to-Houston CSC and the North-to-Houston CSC experienced the greatest amounts of congestion during 2007, although transmission lines on the South-to-Houston interface were upgraded in mid-2007 which greatly reduced the congestion on this interface.

1. Generation Outages and Deratings

Figure 45 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 46 below shows a breakdown of total installed capability for ERCOT on a daily basis during 2007. 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.

Figure 46: Short and Long-Term Deratings of Installed Capability**



* 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 46 shows that long-term outages and other deratings fluctuated between 7 and 22 GW. These outages and deratings reduce the effective resource margins in ERCOT from the levels reported above. Most of these deratings reflect:

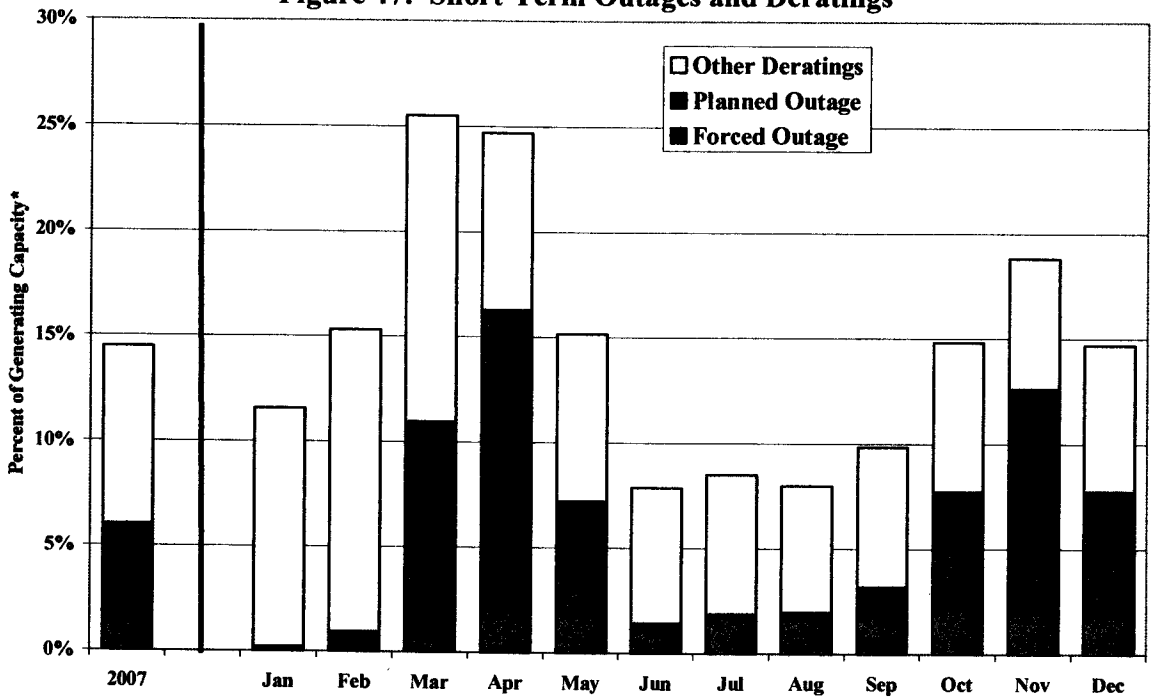
- Cogeneration resources unavailable to serve market load because they are being used to serve self-serve load;
- Resources out-of-service for economic reasons (*e.g.*, mothballed units);
- Output ranges on available generating resources that are not capable of producing up to the full installed capability level (*e.g.*, wind resources); or
- Resources out-of-service for extended periods due to maintenance requirements.

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

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

The next analysis focuses specifically on the short-term forced outages and other short-term deratings. Figure 47 shows the average magnitude of the outages and deratings lasting less than 60 days for the year and for each month during 2007.

Figure 47: Short-Term Outages and Deratings*



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

Figure 47 shows that total short-term deratings and outages were as large as 25 percent of installed capacity in the spring and fall, and dropped below 8 percent for the summer. Most of this fluctuation was due to anticipated planned outages, which ranged as high as 5 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 0.2 percent and 2 percent of total capacity on a monthly average basis during 2007. 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 47 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, wind deratings due to variable wind conditions 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 6 percent on average during the summer in 2007 and as high as 14 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[25] 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.

---

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

To evaluate the commitment of resources in ERCOT, Figure 48 plots the excess capacity in ERCOT during 2007.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 48: Excess On-Line and Quick Start Capacity During Daily Peaks on Weekdays

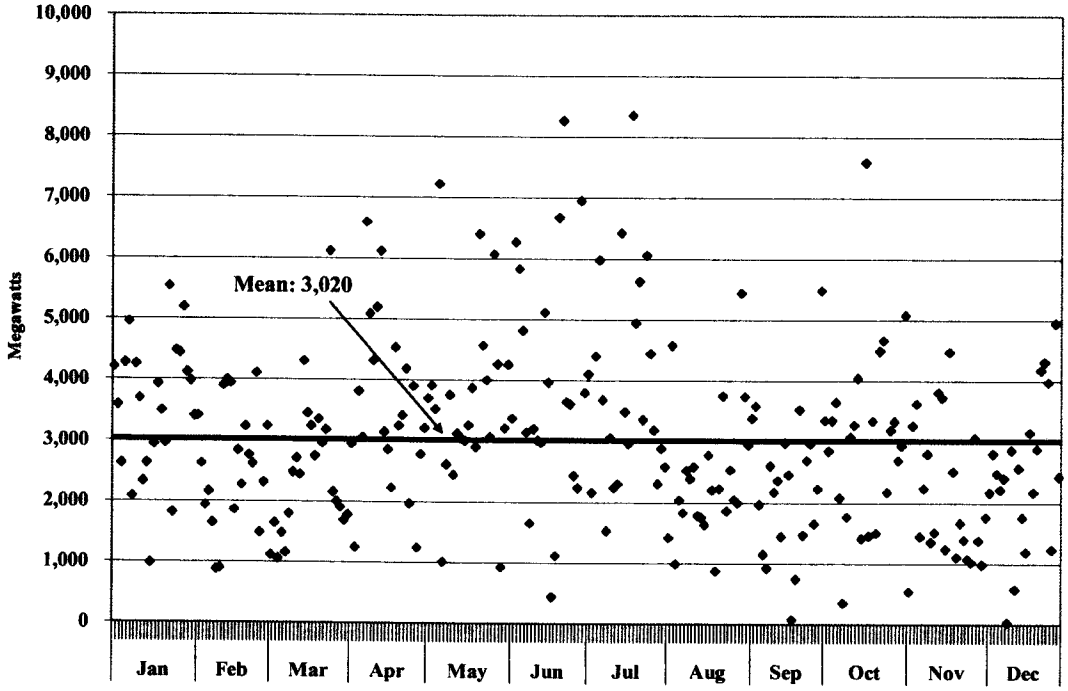


Figure 48 shows that the excess on-line capacity during daily peak hours on weekdays averaged 3,020 MW in 2007, which is approximately 8 percent of the average load in ERCOT. This is at comparable levels as in 2006, with the average daily peak excess on-line capacity being 2,927 MW.

The overall trend in excess on-line capacity also indicates a movement toward more efficient unit commitment across the ERCOT market than 2004 and 2005; 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 promises substantial efficiency improvements in the commitment of generating resources.

C. Demand Response Capability

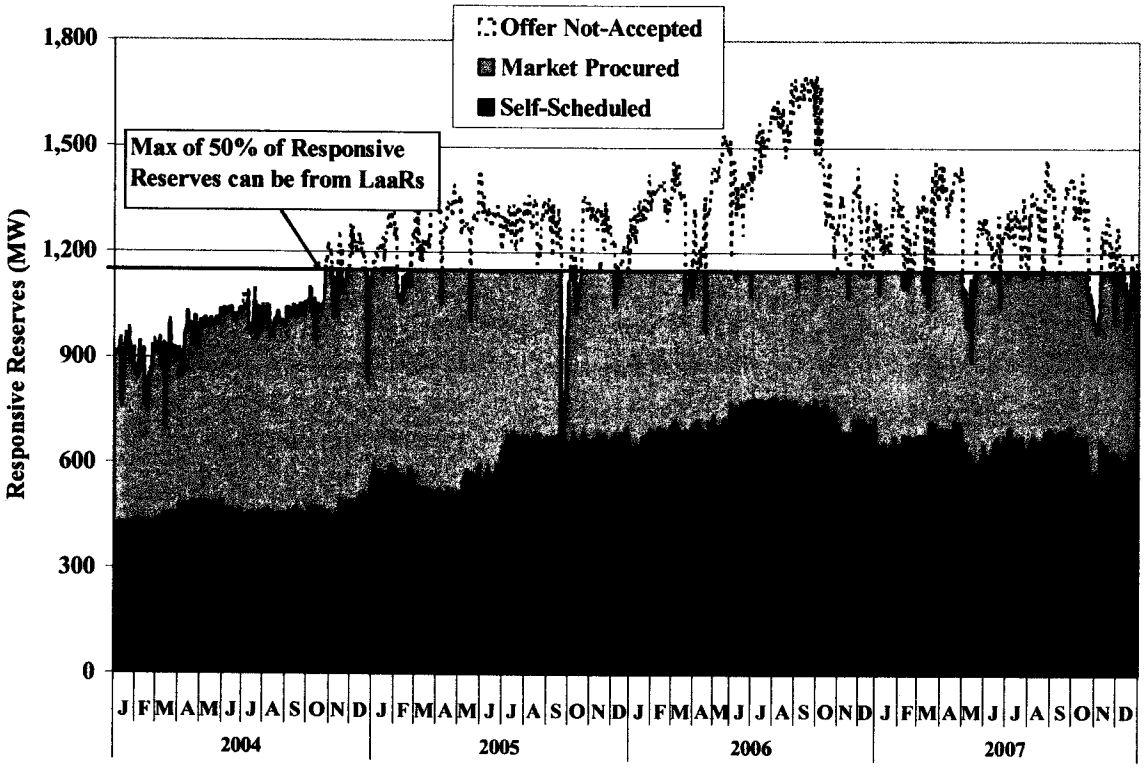
Demand response is a term that broadly refers to actions that can be taken by end users of electricity to reduce load in response to instructions from ERCOT or in response to certain market or system conditions. The ERCOT market allows participants with demand-response capability to provide energy and 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 2007, around 2,050 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. Figure 49 shows the amount of responsive reserves provided from LaaRs on a daily basis in 2007.

Figure 49: Provision of Responsive Reserves by LaaRs
Daily Average



The high level of participation by demand response sets ERCOT apart from other operating electricity markets. Figure 49 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 1,256 MW in 2007. The majority of this increase was procured through self-provision and bilateral agreements rather than the ERCOT administered auction. In 2007, 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.[26] The auction goes up the offer stack until it reaches the 2,300 MW required quantity of

[26] 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

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.

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. 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 reserve pricing and incentives for suppliers, we recommend that ERCOT determine potentially separate prices for responsive reserves by

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.

imposing all supply constraints in the procurement algorithm. 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 was negligible in 2007. 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. 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.

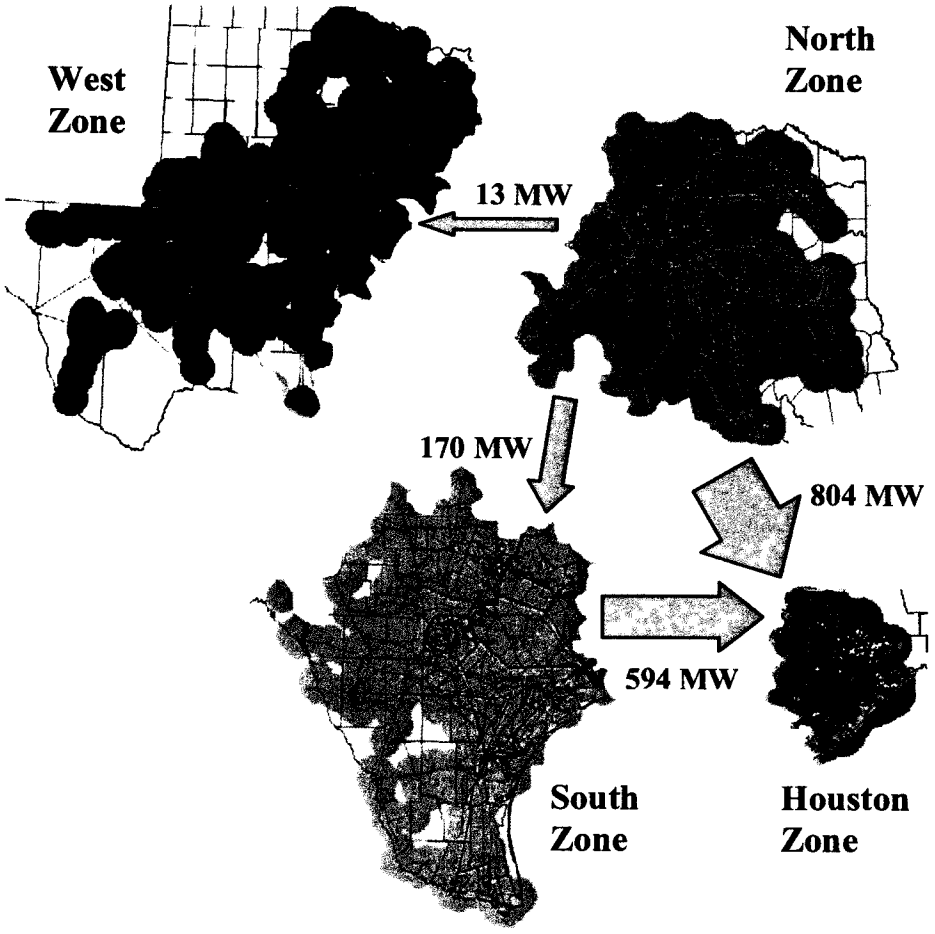
A. Electricity Flows between Zones

In 2007, there were four commercial pricing zones in ERCOT: (a) the North Zone, (b) the West Zone, (c) the South Zone, and (d) the Houston Zone. 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 2006 to 2007. ERCOT operators use the SPD software to economically dispatch balancing energy in each zone to serve load and manage congestion between zones. The SPD model embodies the market rules and requirements documented in the ERCOT protocols.

To manage interzonal congestion, SPD uses a simplified network model with four zone-based locations and five transmission interfaces. These five 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 2007.

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 50 shows the average SPD-modeled flows over CSCs between zones during 2007. A single arrow is shown for the modeled flows of both the North to West and West to North CSCs.

Figure 50: Average SPD-Modeled Flows on Commercially Significant Constraints During All Intervals in 2007



Note: In the figure above, CSC flows are averaged taking the direction into account. So one arrow shows the average flow for the North-to-West CSC was 13 MW, which is equivalent to saying that the average for the West-to-North CSC was *negative* 13 MW.

Figure 50 shows the four ERCOT geographic zones as well as the five 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 North to Houston interface, and (e) the North to West interface. Based on SPD modeled flows, Houston is a significant importer while the North and South Zones export significant amounts of power.

The most important simplifying assumption underlying the zonal model is that all generators in a zone have the same effect on the flows over the CSC, or the same generation shift factor ("GSF")[27] 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. 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

CSC 2007	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	-13	-29	*-15*	-133	*-104*
South-North	-170	-154	*17*	-75	*78*
South-Houston	594	592	*-2*	834	*242*
North-Houston	804	794	*-10*	650	*-144*

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

[27] 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.

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 and North to Houston, 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 242 MW and reduced the calculated flows on the North to Houston CSC by 144 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 2007, the five 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 2007.

Table 3: Actual Net Imports vs. SPD-Calculated Flows on CSCs 2003 to 2007

Year	Zone	Actual Net Imports	SPD Flows on CSCs
2003	Houston	1,796	565
	North	-507	191
	South	-1,213	-702
	West	-76	-54
2004	Houston	2,479	1,265
	North	867	264
	NorthEast	-2,116	-858
	South	-1,531	-800
	West	304	129
2005	Houston	2,596	1,247
	North	660	164
	NorthEast	-2,138	-845
	South	-1,501	-728
	West	386	162
2006	Houston	3,434	1,744
	North	462	20
	NorthEast	-2,334	-974
	South	-1,741	-870
	West	180	79
2007	Houston	3,264	1,398
	North	-2,019	-1,001
	South	-1,319	-764
	West	74	13

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 2007. 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 and then stayed about the same level in 2007.[28] The West Zone shifted from being a net exporter in 2003 to importing substantial quantities from 2004 to 2007, with the average import levels dropping by about 58 percent in 2007 compared to 2006. From 2003 to 2007, net exports increased from the North zone compared with the combined area of the North and Northeast zones from 2004 to 2006. Net exports from the South zone increased from 2003 to 2006, and dropped about 24 percent in 2007. In every case, the SPD-calculated flows on CSCs were significantly less than the actual interchange.

28 The North to Houston CSC was added in 2004.

B. Interzonal Congestion

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

Figure 51 shows the average SPD-calculated flows between the four 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 51: Average SPD-Modeled Flows on Commercially Significant Constraints During Transmission Constrained Intervals in 2007

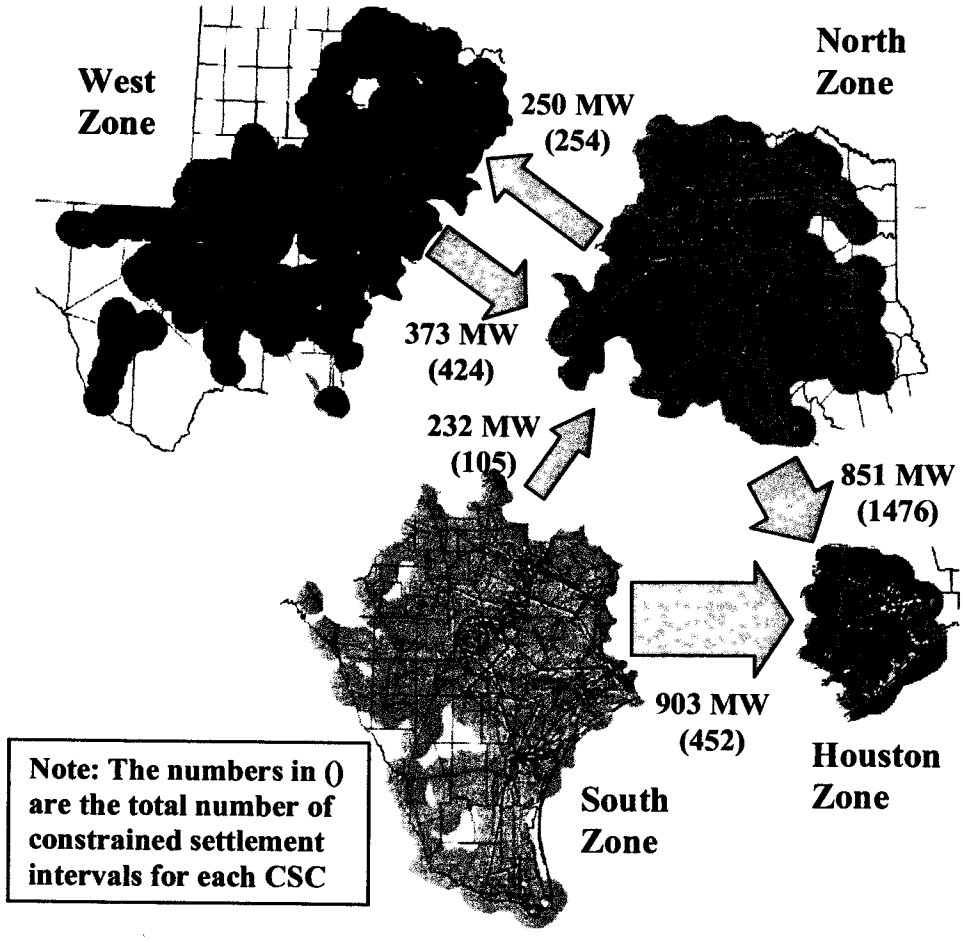


Figure 51 shows that inter-zonal congestion was most significant on the North to Houston CSC which exhibited SPD-calculated flows averaging 851 MW during 1,476 constrained intervals in 2007. Congestion was also significant on the South to Houston and West to North CSCs.

1. Congestion Rights in 2007

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 is 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 2007. Figure 52 shows the average number of TCRs and PCRs that were allocated for each of the CSCs in 2007, as well as the average SPD-modeled flows during the constrained intervals.

Figure 52: Transmission Rights vs. Real-Time SPD-Calculated Flows Constrained Intervals

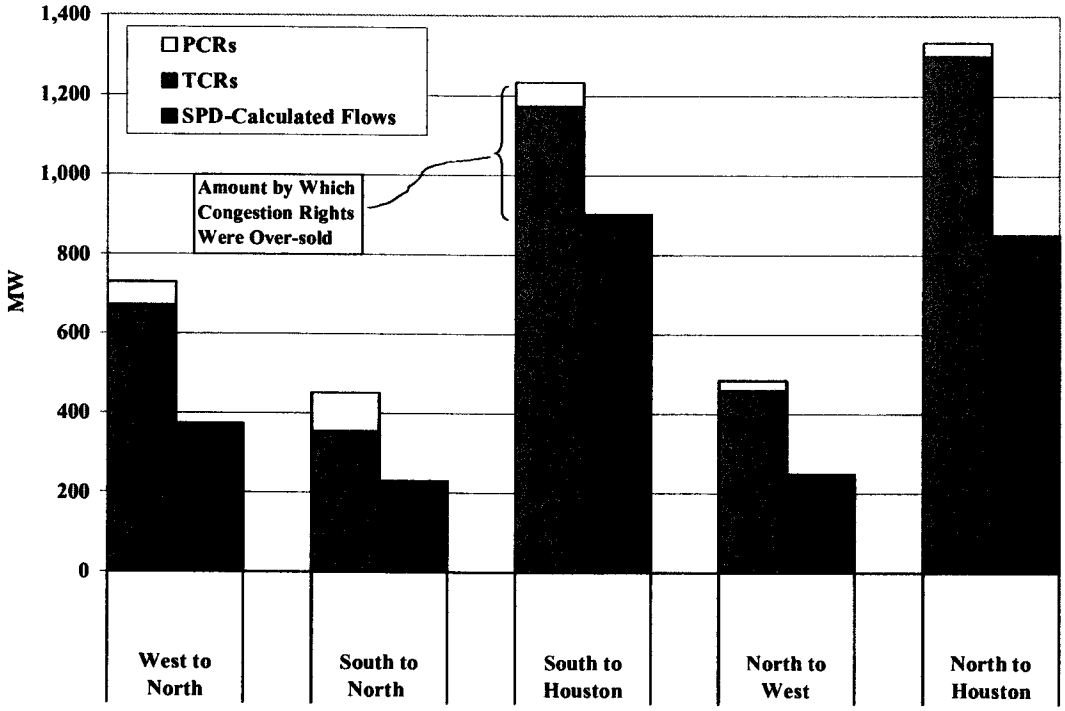


Figure 52 shows that total congestion rights (the sum of PCRs and TCRs) on all the interfaces exceeded the average real-time SPD-calculated flows during constrained intervals. These results indicate that the congestion rights were oversold in relation to the SPD-calculated limits for some CSCs. For instance, congestion rights for the North to Houston CSC were oversold by an average of 482 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 52, we compare the SPD-calculated flows and congestion rights quantities for each of the constrained intervals by CSC. In addition to the observation of SPD flow versus the congestion rights, we also present the relationship between actual flows versus the actual CSC limits. Over-constraining the CSC limit will cause unnecessary congestion costs and will distort balancing energy market prices, which are undesirable for an efficient market. Under-constraining the CSC limit will cause reliability issues. Although exact matching of the actual flow with the actual physical limit is ideal, fluctuations of the actual flows around the physical limit are expected due to the simplifying assumptions of the zonal market model. However, significant divergence is not desirable.

2. Congestion on South to North CSC

Figure 53 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 2007. Because only congested intervals are shown, some months will have significantly more observations than other months. Although some congestion occurred in every month, the month of November accounted for 28 percent of all constrained intervals during 2007 due to a number of planned transmission outages.

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 53 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 53: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to North

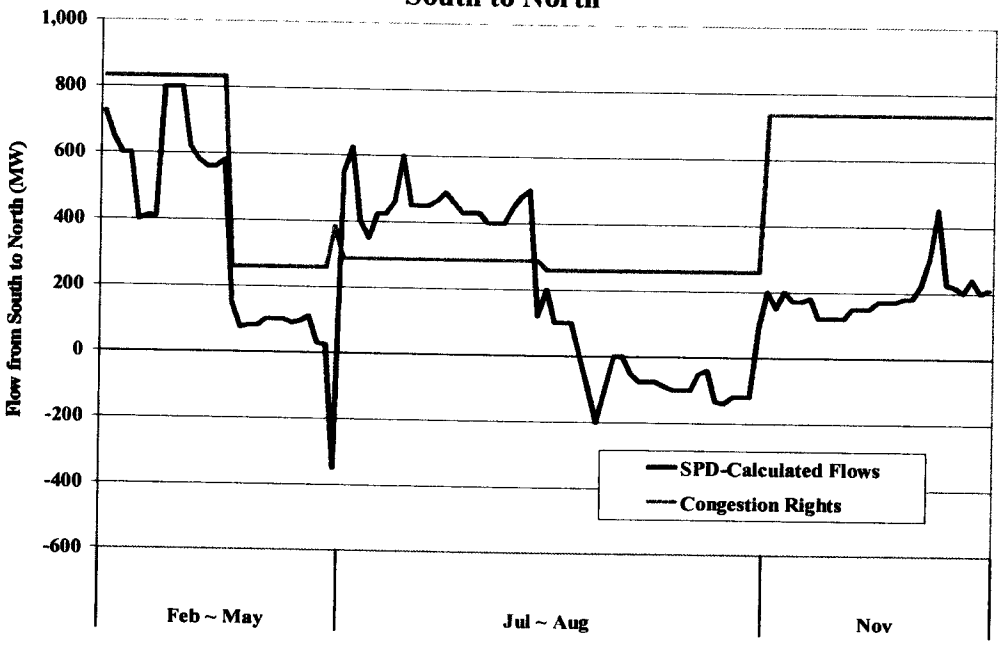


Figure 53 indicates that the quantity of outstanding congestion rights fluctuated considerably during 2007. In November, more than 700 MW of rights for the South to North CSC were available, whereas for March, July and August, less than 300 MW of congestion rights were allocated for the South to North CSC in 2007. 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.

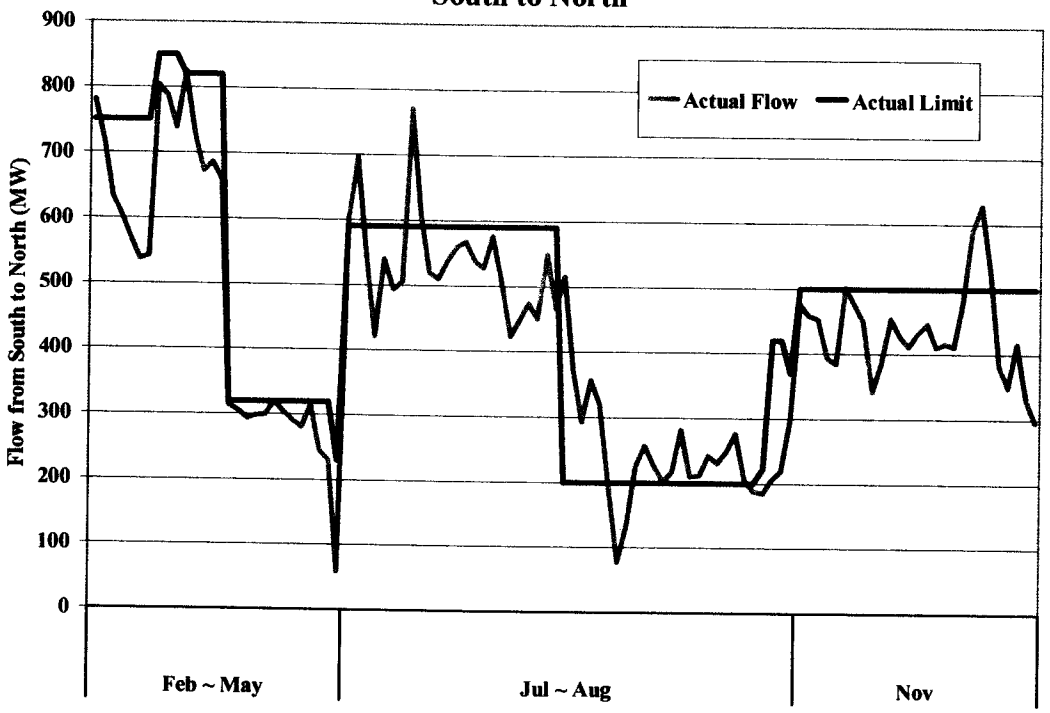
For the South to North CSC, the congestion rights were above and below SPD flows during different months for the congested intervals in 2007. The figure shows ten constrained intervals when the SPD-calculated flows were *negative* at times during May and August.

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.

In 2006, the South-to-North CSC congested 583 times with an average flow of 582 MW, while in 2007, it congested only 105 times with an average flow of 232 MW. Along with the reduction in South-to-North congestion, there was an increase in congestion from North-to-South in 2007. Because there was not a North-to-South CSC defined for 2007, this congestion was managed with local congestion management. However, in response to these changing congestion patterns, a new CSC was added for the North-to-South interface for 2008.

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, which would cause unnecessary congestion in some extreme cases. The following figure presents the South to North actual flow versus the actual South to North limit.

Figure 54: Actual Flows versus Physical Limits during Congestion Intervals South to North



The South to North CSC experienced 105 intervals of congestion during 2007. During the congestion intervals, the actual flow amount over the CSC was less than the actual physical limit by an average of 80 MW. Because of the long times between the dispatch decisions and the operating interval, as well as the simplifying assumptions of the zonal model, the tendency of ERCOT operations in the zonal model is to operate more conservatively and over-constrain CSCs. As also shown in the figures in the following subsections, this was true for all of the CSCs in 2007. The implementation of the nodal market will improve the efficiency of the management of these constraints by providing more frequent re-dispatch that utilizes data that is more reflective of current operating conditions, and by relying upon a commercial model that is consistent with the operational reality.

3. Congestion on South to Houston CSC

This interface experienced 452 constrained intervals, reduced significantly from 2006, when it congested 1,001 times. The most congestion occurred in May and June due to transmission outages associated with the construction of new transmission lines that effectively relieved the congestion on this interface for the remainder of the year, with the exception of October when

transmission maintenance outages occurred. In the months with significant congestion, SPD flows averaged between 1,020 and 1,156 MW and the congestion rights were about the average level of SPD flows. However, congestion rights were above the SPD flow levels in the months of February through April as well as in October. Figure 55 shows the comparison between actual flow and the congestion rights quantities.

Figure 55: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals South to Houston

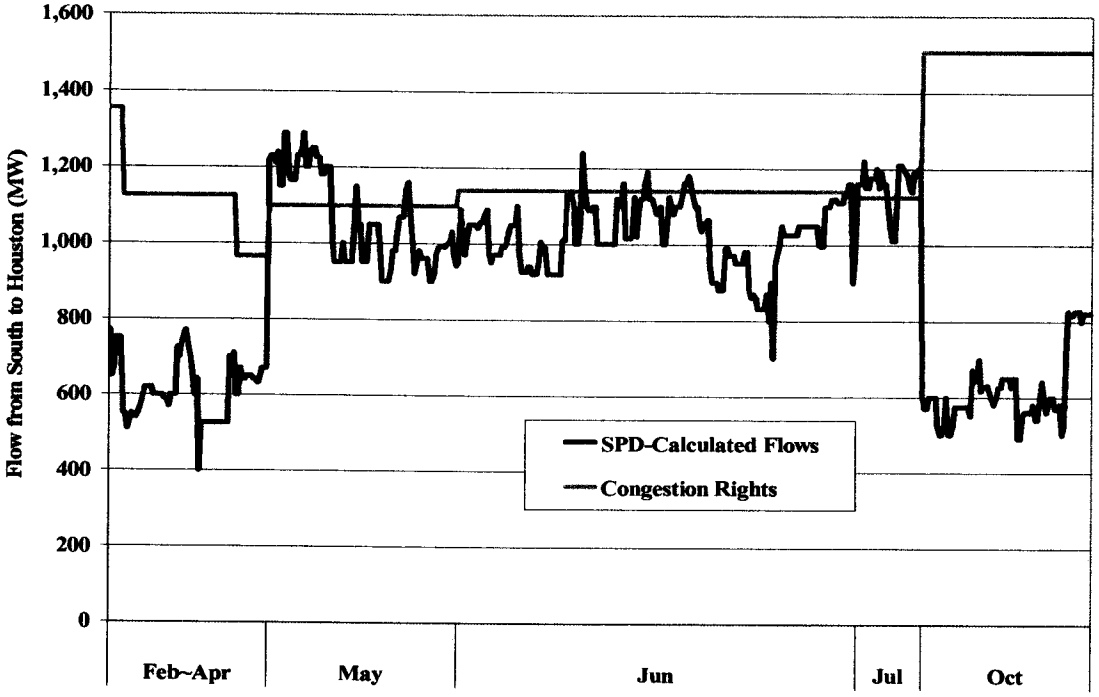


Figure 56: Actual Flows versus Physical Limits During Congestion Intervals South to Houston

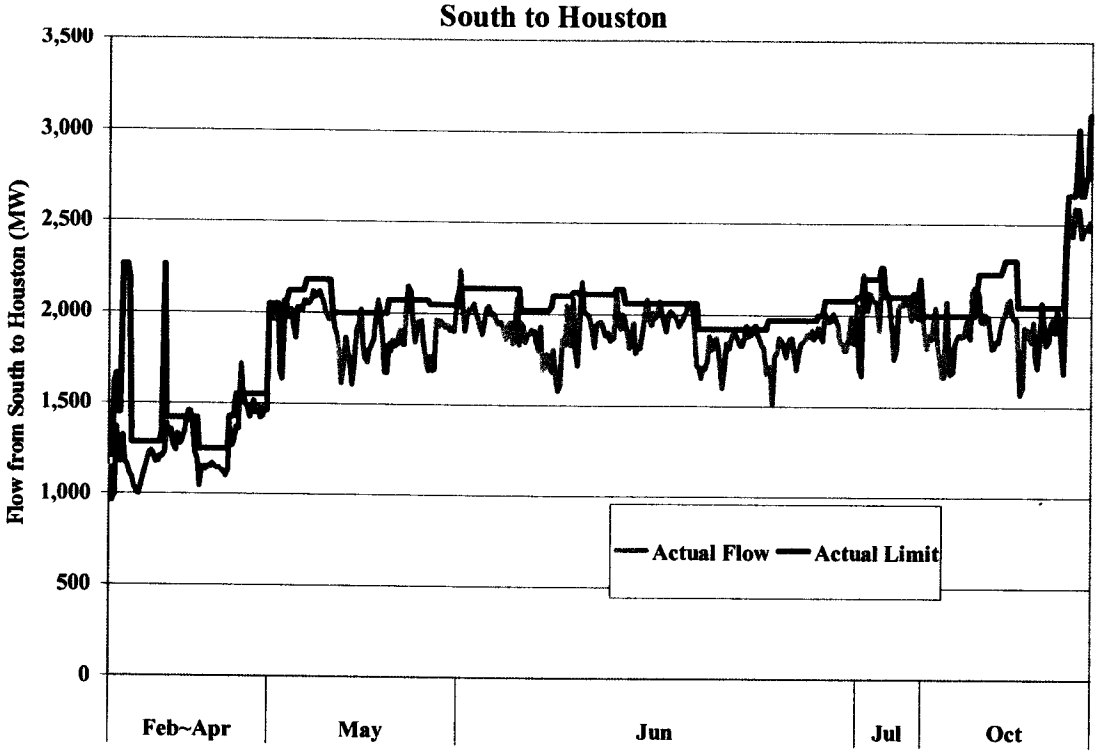


Figure 56 compares the actual flow with the actual limit for the South to Houston CSC. During the congestion intervals, the actual flow over the CSC was less than the physical limit by an average of 169 MW.

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. The congestion rights were almost in line with the average SPD flows during the months of June to October. However, the congestion rights were above the SPD flow levels during the months of April and November and below the SPD flow levels during the month of May. In November, the number of congestion rights allocated were above the average SPD flow levels during congestion periods by 1,003 MW due to planned transmission outages that were not accounted for at the time of the TCR auction. The frequency of transmission constraints rose dramatically in November in conjunction with the increase of rights allocated. In 2007, this interface became the most congested interface with congestion occurring in 1,476 intervals, with a significant portion of the congestion occurring in November.

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

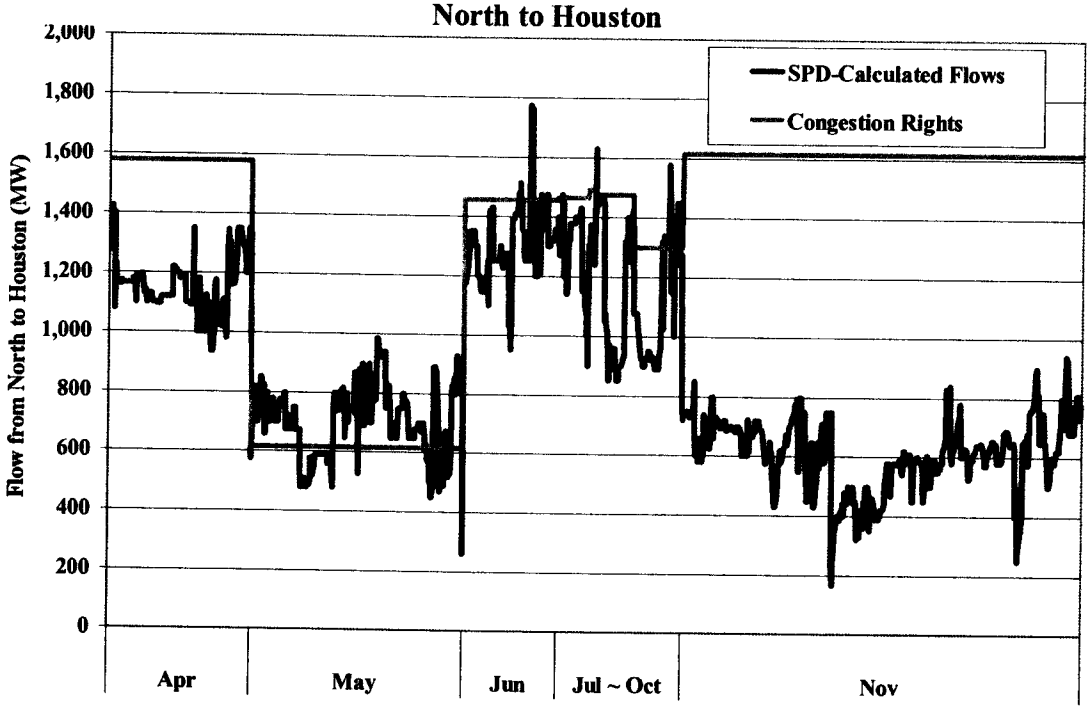


Figure 58: Actual Flows versus Physical Limits during Congestion Intervals North to Houston

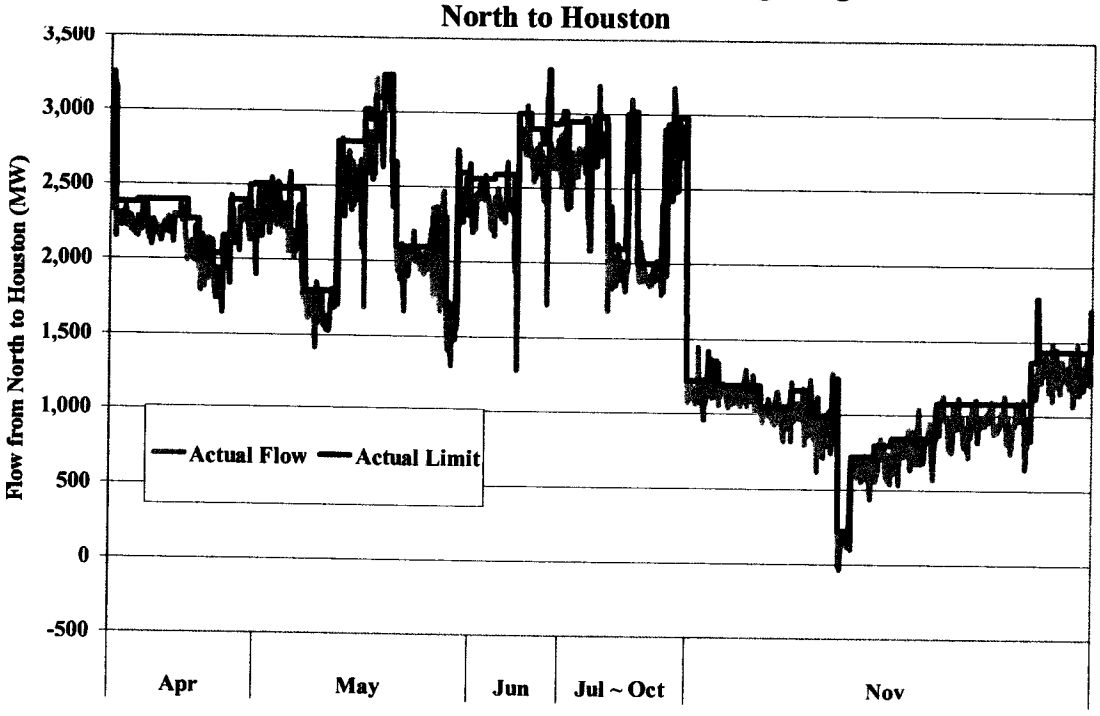


Figure 58 compares the actual flow with the actual limit for the North to Houston CSC. During the congestion intervals, the actual flow over the CSC was less than the physical limit by an average of 167 MW.

5. Congestion on North to West CSC

This CSC was congested most frequently during the winter months with approximately 39 percent of constrained intervals in November to December. Congestion rights were above the SPD flows in the months of January through May and also August through December. Although the number of congestion rights allocated for this interface varied from 178 to 823 MW over the year, the SPD flows averaged just 250 MW during constrained intervals.

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

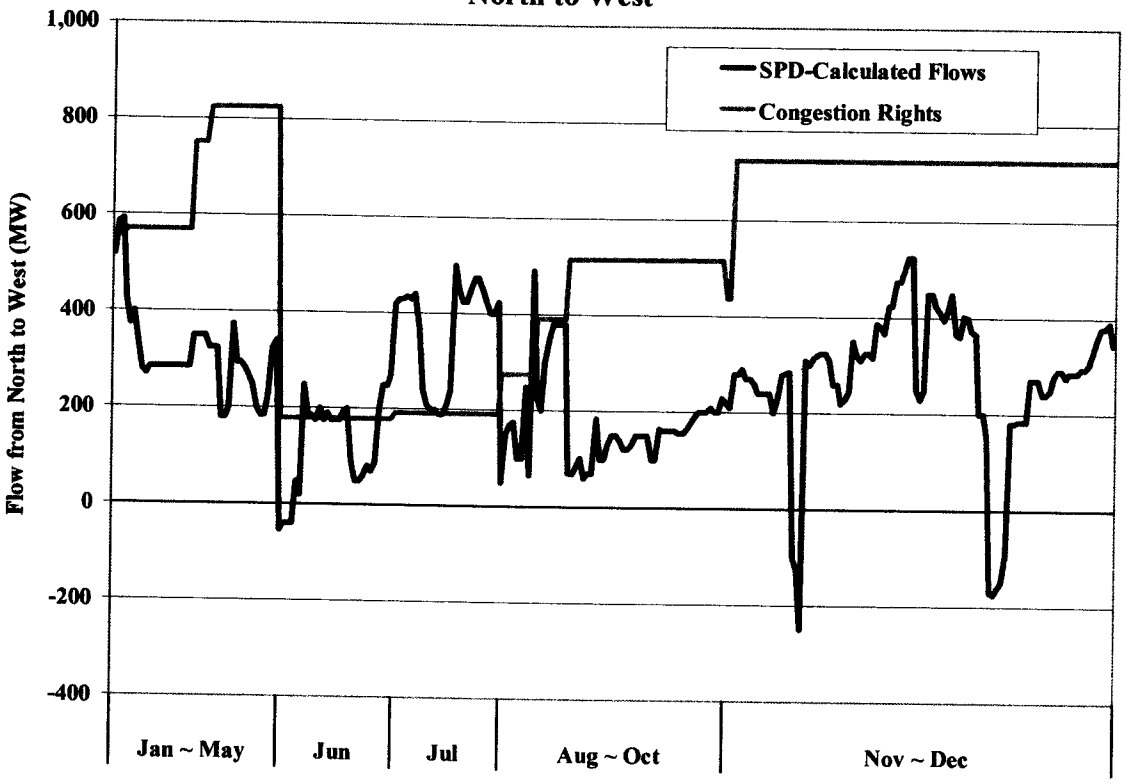


Figure 60: Actual Flows versus Physical Limits during Congestion Intervals North to West

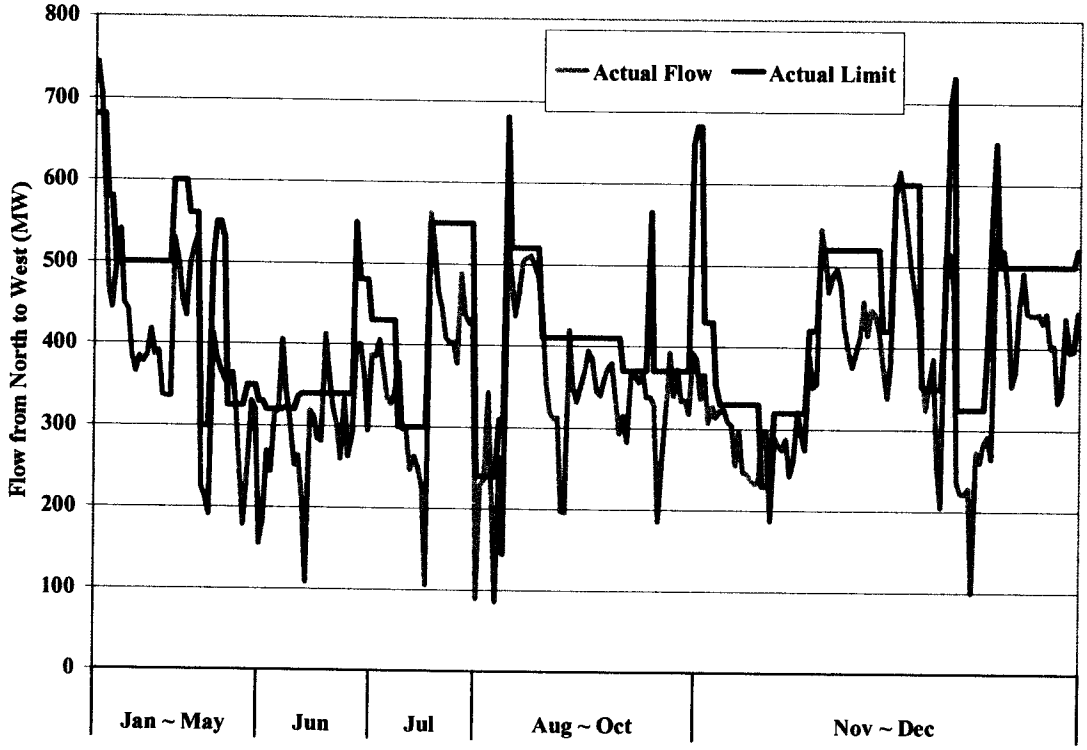


Figure 60 compares the actual flow with the actual limit for the North to West CSC. During the congestion intervals, the actual flow over the CSC was less than the physical limit by an average of 78 MW.

6. Congestion on West to North CSC

This CSC was congested in 424 intervals during 2007, much more than the congestion frequency in 2006 of 48 intervals. Most of the increase occurred in the last quarter of 2007, and is associated with the significant increases in wind generation in the West Zone during this time period. Different from other CSCs, the TCRs allocated were almost always higher than the actual SPD flow during congestion intervals. The average SPD flow during congestion intervals was 373 MW and the average TCR sold on the CSC was 795 MW. The main reason for the difference is due to planned and unplanned transmission outages that are not accounted for in the TCR auctions that significantly reduce the real-time transfer capability of the CSC. In addition, at times there are dynamic stability limits for West-to-North transfers that may result in limits that are much lower than the transfer capability used to determine TCR auction quantities.

Figure 61: Congestion Rights Allocated vs. SPD Flows during Constrained Intervals West to North

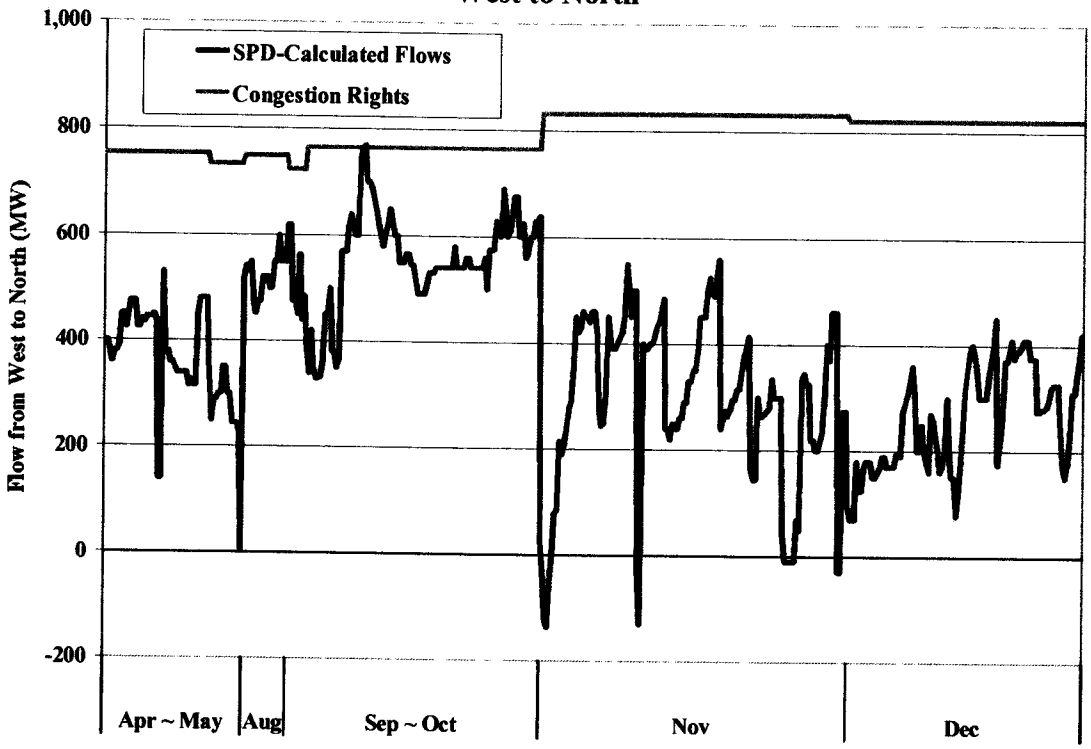


Figure 62: Actual Flows versus Physical Limits during Congestion Intervals West to North

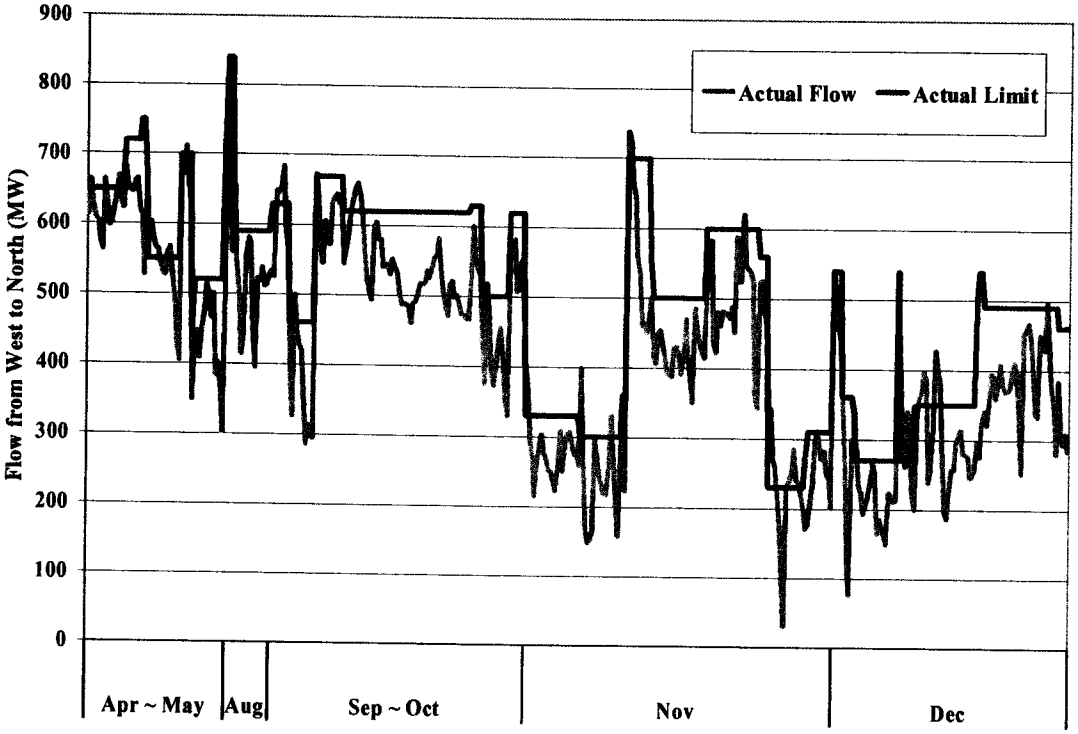


Figure 62 compares the actual flow with the actual limit for the West to North CSC. During the congestion intervals, the actual flow over the CSC was less than the physical limit by an average of 81 MW.

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.[29] Since the monthly studies tend to more accurately reflect conditions that will prevail in the coming month, the monthly designations tend to more closely reflect actual transmission limits.

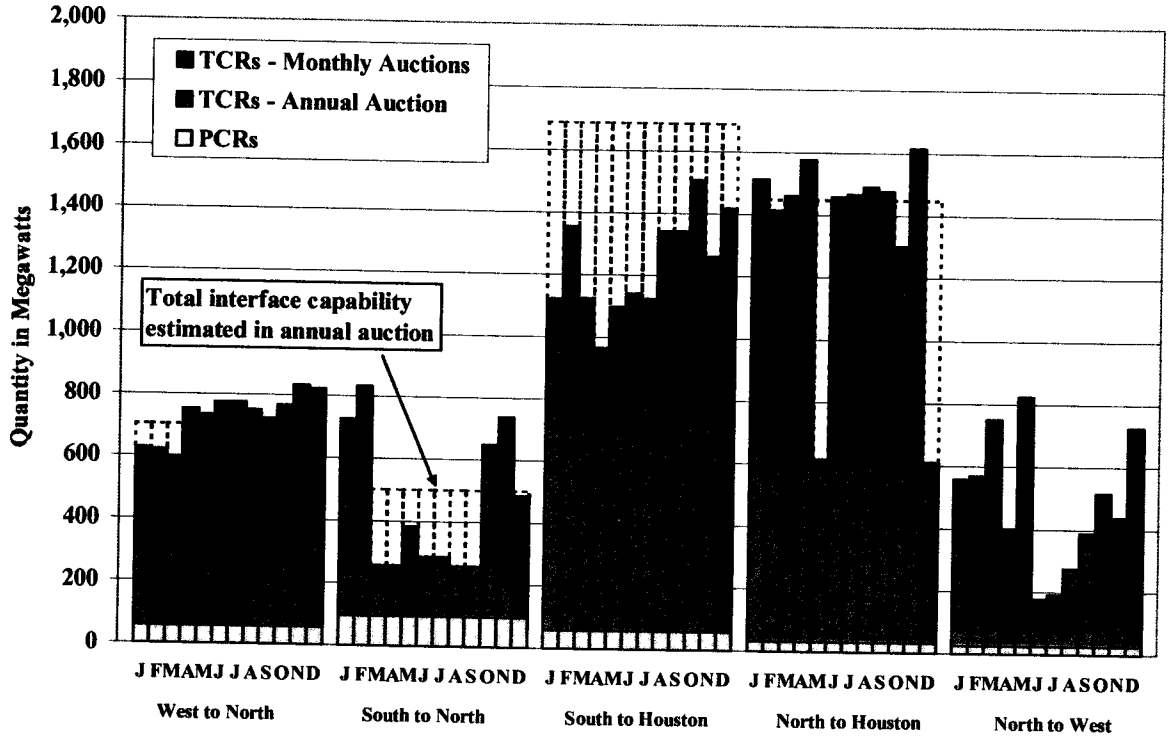
However, the summer monthly studies used to designate the TCRs do not always accurately reflect transmission conditions that can arise in real-time. This happens for two main reasons. First, transmission and generation outages can occur unexpectedly and significantly reduce the transfer capability of a CSC, and even planned transmission outages may not be known to ERCOT when the summer studies are conducted. 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 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 63 shows the quantity of each category of congestion rights for each month during 2007. The quantities of PCRs and annual TCRs are constant across months and

[29] 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.

were determined before the beginning of 2007, while monthly TCR quantities can be adjusted monthly.

Figure 63: Quantity of Congestion Rights Sold by Type 2007

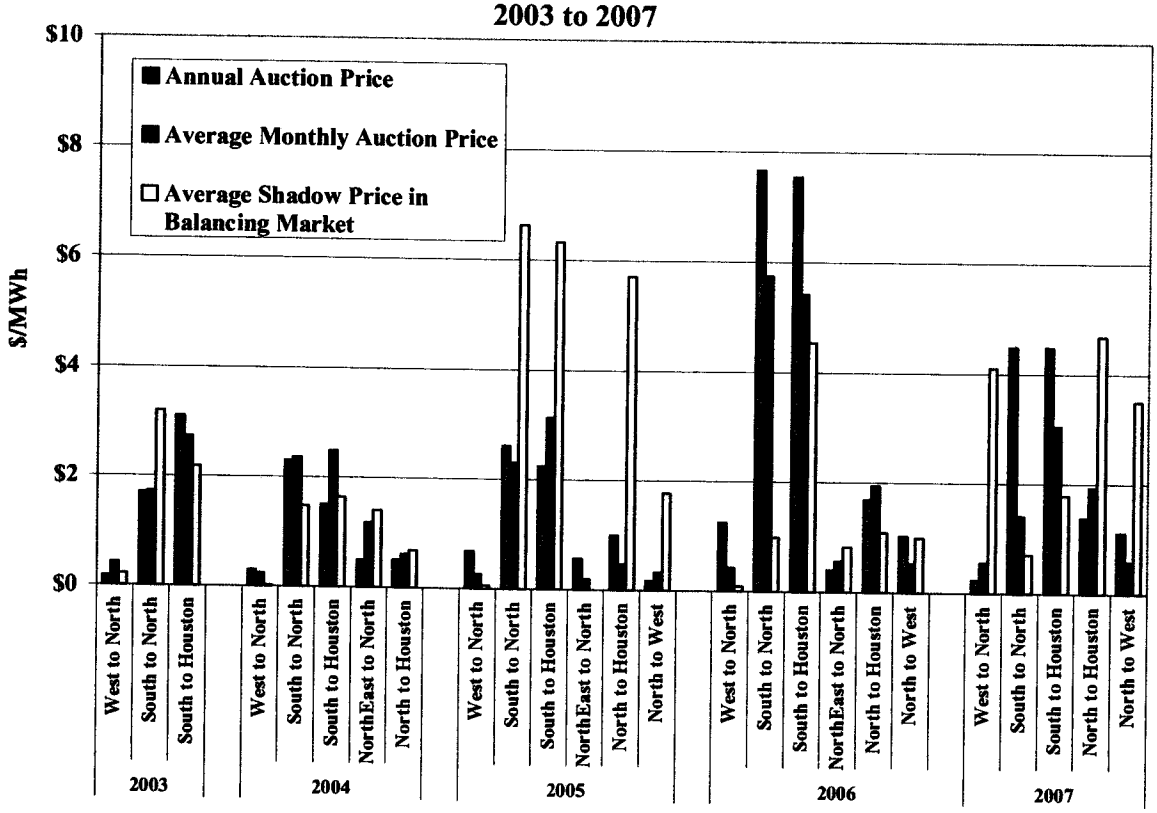


When the monthly planning studies indicate changes from the summer study, revisions are often made to the estimated transmission capability. Therefore, the auctioned congestion rights may increase or decrease relative to the amount estimated in the summer study. The shadow boxes in the figure represent the capability estimated in the summer study that is not ultimately sold in the monthly auction. When there is no shadow box in Figure 63, 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, South to Houston 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, the South to North TCRs were not even auctioned during four of 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 64 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.

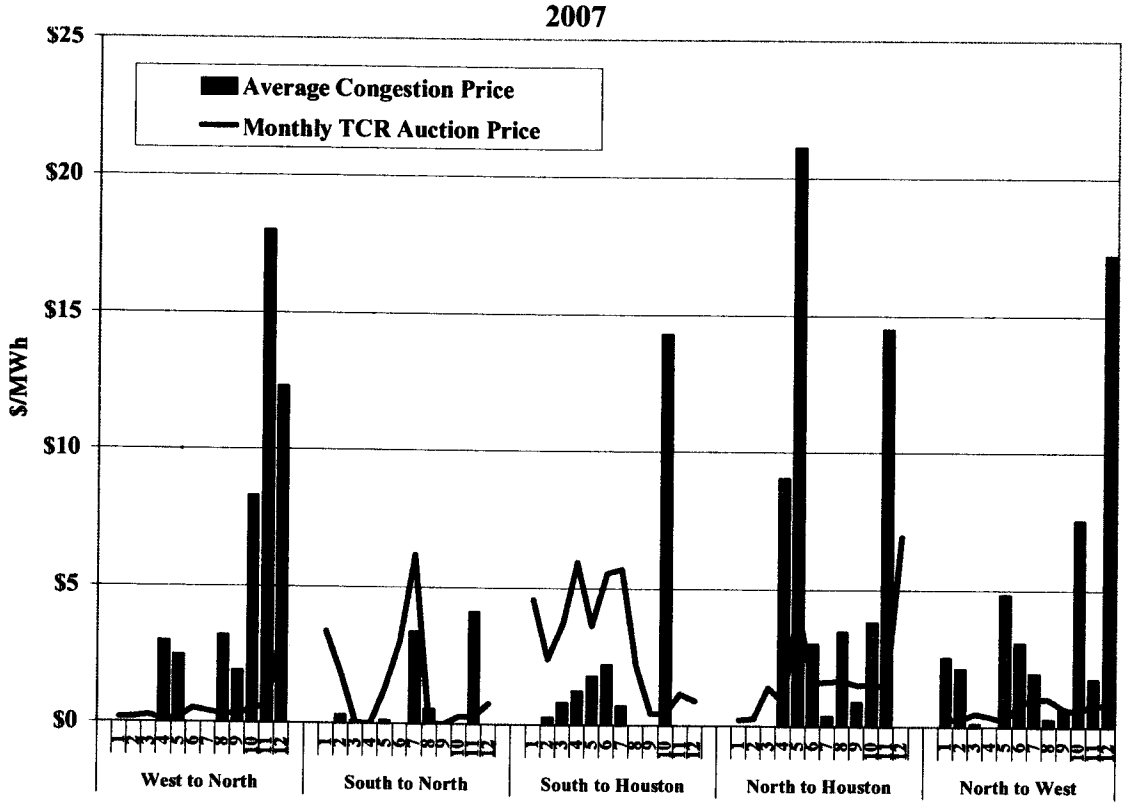
Figure 64: TCR Auction Prices versus Balancing Market Congestion Prices 2003 to 2007



This figure shows that there is a tendency for the TCRs to settle at prices that are closer to the previous years' value, but that real-time congestion prices often diverge significantly from auction prices. This suggests that participants are not able to forecast annual interzonal congestion costs and accurately value the TCRs in the annual auction, and instead rely more upon historical market outcomes.

Figure 65 compares monthly TCR auction prices with monthly average real-time CSC shadow prices from SPD for 2007. The TCR auction prices are expressed in dollars per MWh.

Figure 65: Monthly TCR Auction Price and Average Congestion Value 2007



The TCR price trends for North to Houston CSCs correlated well with the actual congestion prices, although the TCR prices for this CSC are far below the congestion prices. Overall, market participants did a poor job predicting fluctuations in congestion during 2007, particularly on the South to Houston interfaces. For South to Houston interfaces, there was one month when balancing market congestion spiked when balancing prices far exceeding the TCR prices.

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.[30] 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 66 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.

30 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 66: TCR Auction Revenues, Credit Payments, and Congestion Rent[31]
2003 to 2007

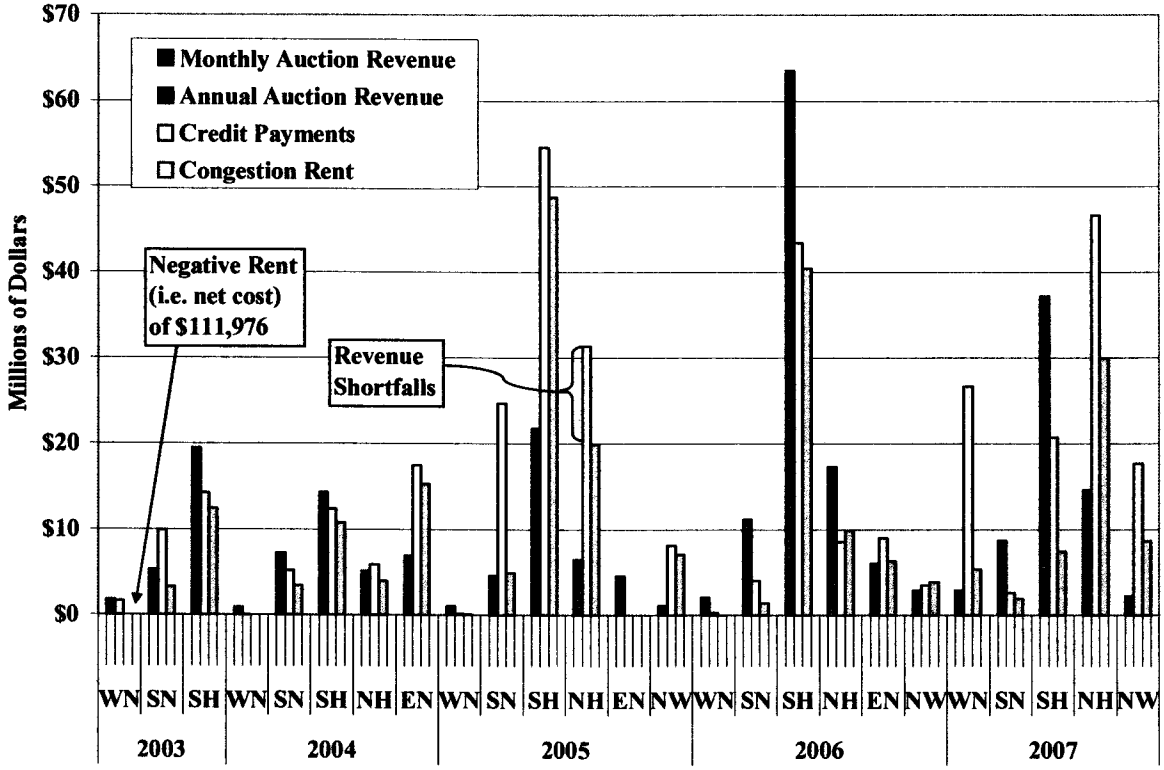


Figure 66 shows that in 2004, the auction revenues were consistent with credit payments for the three CSC that existed in 2003. This appeared to be due to market participant basing their valuations of the TCRs on their value in the prior year. The auction revenues for the North to Houston CSC, which was added for the first time in 2004, were quite close to credit payments. However, market participants substantially under-valued congestion on the Northeast to North interface, which was also new in 2004.

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

31 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.

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 66, 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.

In 2007, the South to North and South to Houston interfaces exhibited similar pattern as in 2006, where the auction revenue exceeded credit payments. In contrast, the West to North, North to West and the North to Houston interface show signification higher credit payments than auction revenue, while there are still revenue short falls on those three interfaces since credit payments also exceeded congestion rent.

Figure 66 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 2004 when congestion rents covered 81 percent of payments to TCR holders. 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. In 2007, however, congestion rents only covered 47 percent of payments to TCR holders. 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.[32] These shortfalls are included in the

[32] 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

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.

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

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.

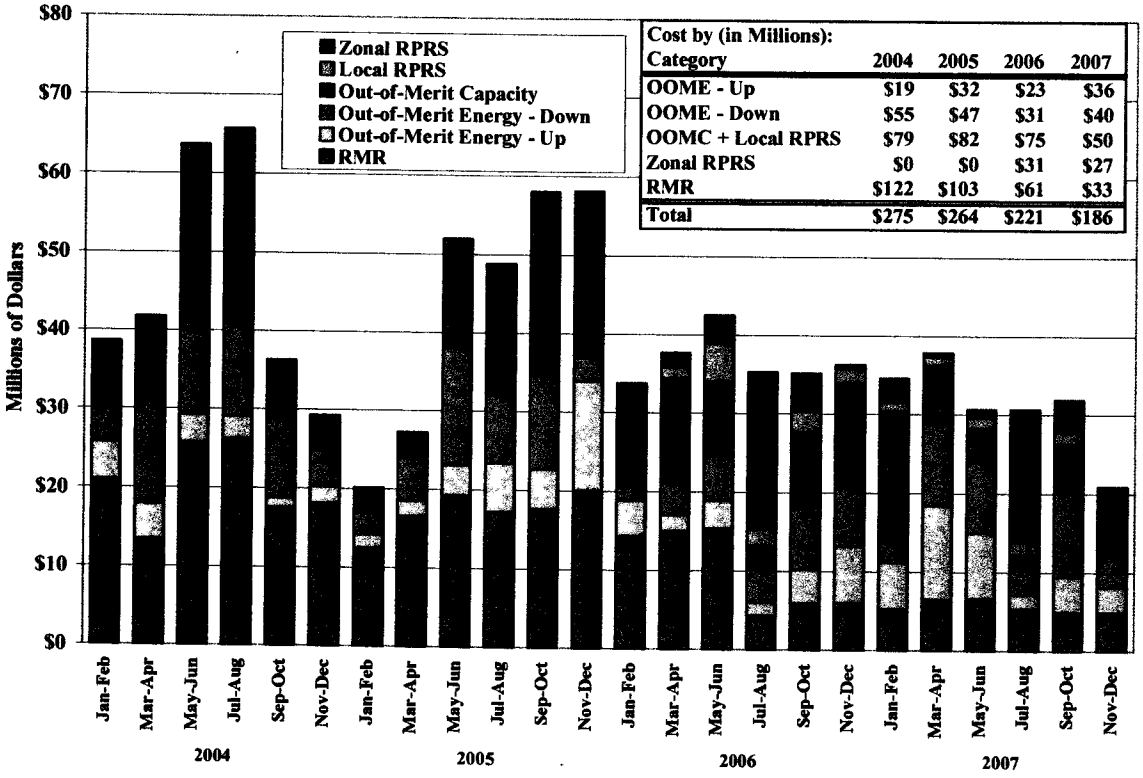
and the actual output resulting from the OOME instruction from ERCOT. The payment per MWh for OOME is a pre-determined amount specified in the ERCOT Protocols based on the type and size of the unit, the natural gas price, and the balancing energy price. The net payment to a resource receiving an OOME 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. In 2007, there were only three RMR units (Laredo units 1, 2 and 3). Units contracted to provide RMR service to ERCOT are compensated for start-up costs, energy costs, and are also paid a standby fee. Figure 67 shows each of the four categories of uplift costs from 2004 to 2007.

Figure 67: Expenses for Out-of-Merit Capacity and Energy 2004 to 2007



Cost by (in Millions): Category	2004	2005	2006	2007
OOME - Up	$19	$32	$23	$36
OOME - Down	$55	$47	$31	$40
OOMC + Local RPRS	$79	$82	$75	$50
Zonal RPRS	$0	$0	$31	$27
RMR	$122	$103	$61	$33
Total	$275	$264	$221	$186

The results in Figure 67 show that overall uplift costs for RMR units, OOME units, OOMC/Local RPRS and Zonal RPRS[33] units decreased in 2007 from the 2006 level. The costs decreased by $74 million in 2006 from $264 million to $221 million. The cost further decreased by $35 million in 2007. As previously noted, there were substantial reductions to RMR cost due to the expiration of RMR agreements, which accounts for $28 million of the $35 million decrease from 2006 to 2007. Total OOME Up and OOME Down costs increased from $54 million in 2006 to $76 million in 2007. A sizable portion of this increase can be attributed to the management of North-to-South congestion in 2007 during which there was not a CSC defined for this interface. Unit commitment cost decreased in 2007 by $29 million from 2006. Notably, zonal RPRS costs for system adequacy were the highest in the peak system demand months of

33 Zonal RPRS for system adequacy is deployed at the second stage of the RPRS run, which is affected by the deployment at the first stage of the RPRS run, or the local RPRS deployment. Because ERCOT Protocols allocate the costs of local and zonal RPRS in the same manner, we have included both as local congestion costs.