POTOMAC ECONOMICS

2004 ASSESSMENT OF THE OPERATION OF THE ERCOT WHOLESALE ELECTRICITY MARKETS

POTOMAC ECONOMICS, LTD.

Advisor to the Market Oversight Division
Public Utility Commission of Texas

November 2004

TABLE OF CONTENTS

Executive Summary 1

I. Interzonal Congestion Management 30

A. Background 30
B. Summary of Interzonal Congestion 32
C. Zonal Congestion Management and Generation Shift Factors 34
D. Assessment of Zonal Shift Factors 42
E. Accuracy of Modeled Flows 54
F. Redispatch Analysis 61
G. Interzonal Congestion Conclusions and Recommendations 65

II. Local Congestion Management 67

A. Summary of Congestion Costs during the Study Period 67
B. Comparison of Local and Interzonal Congestion 76
C. Multi-Step Congestion Management Process 82
D. Impact of Local Congestion on Balancing Prices 85
E. Local Congestion Conclusions 88

III. Load Forecasting 90

A. Day-ahead forecasting 90
B. Real-time forecasting 93

IV. Real-Time Market Operations 98

A. Scheduling and Balancing energy market Outcomes 99
B. Real-Time Operations and System Control 104
B. Portfolio Ramp Constraints in SPD 127
C. QSE Provision of Reserves 133

LIST OF FIGURES

Figure 1: ERCOT Zones and CSCs 31
Figure 2: Daily Percentage of Intervals Congested by CSC 33
Figure 3: Total Redispatch Impact by CSC 37
Figure 4: OC 0 vs. OC 1 Limits During Congested Intervals: South-to-North CSC 38
Figure 5: OC 0 vs. OC 1 Limits During Congested Intervals: South-to-Houston CSC 39
Figure 6: OC 0 vs. OC 1 Limits During Congested Intervals: North-to-Houston CSC 40
Figure 7: OC 0 vs. OC 1 Limits During Congested Intervals: Northeast-to-North CSC ... 41
Figure 8: Distribution of Resource-Specific GSFs by Zone West-to-North CSC 43
Figure 9: Distribution of Resource-Specific GSFs by Zone South-to-North CSC 44
Figure 10: Distribution of Resource-Specific GSFs by Zone South-to-Houston CSC 46
Figure 11: Distribution of Resource-Specific GSFs by Zone North-to-Houston CSC 47
Figure 12: Distribution of Resource-Specific GSFs by Zone Northeast-to-North CSC 48
Figure 13: Actual Flows vs. SPD Flows on the West-to-North Interface 55
Figure 14: Actual Flows vs. SPD Flows on the South-to-North Interface 57
Figure 15: Actual Flows vs. SPD Flows on the South-to-Houston Interface 58
Figure 16: Actual Flows vs. SPD Flows on the North-to-Houston Interface 59
Figure 17: Actual Flows vs. SPD Flows on the Northeast-to-North Interface 60
Figure 18: Expenses for Out-of-Merit Commitment and Dispatch 69
Figure 19: Expenses for OOMC and RMR by Region 70
Figure 20: Expenses of Out-of-Merit Dispatch by Region 72
Figure 21: Frequency of Interzonal and Local Congestion 77
Figure 22: Payments to TCR Holders vs. Local Congestion Payments 79
Figure 23: Comparison of Market-Based and Local Congestion Costs 80
Figure 24: Forecast Error at Daily Peak 91
Figure 25: Day-Ahead Load Forecast by Hour 92
Figure 26: SPD Load vs. Actual Load 94
Figure 27: Average Deviation With and Without Offset Adjustment 96
Figure 28: Final Schedules during Ramping-Up Hours 99
Figure 29: Final Schedules during Ramping-Down Hours 100
Figure 30: Balancing Energy Prices and Volumes Ramping-Up Hours 102
Figure 31: Balancing Energy Prices and Volumes Ramping-Down Hours 102
Figure 32: Regulation Need by Time of Day 105
Figure 33: Percent of Intervals with Large Need for Regulation 107
Figure 34: Regulation Need from Actual Offset vs. Alternative Offset 109
Figure 35: Average ACE by Minute 111
Figure 36: Percent of Intervals with ACE Higher than 450 MW 113
Figure 37: Schedule Control Error and Load Deviations 114
Figure 38: SPD Load vs. Actual Load during Ramping Hours 116
Figure 39: SCE and Load Deviations during Ramping Intervals 117
Figure 40: SCE and Load Deviations in Periods with Large Negative Regulation Need 119
Figure 41: SCE and Load Deviations in Periods with Large Positive Regulation Need .. 120
Figure 42: Total SCE for Large and Small QSEs 121
Figure 43: Average SCE by QSE during Afternoon Hours 122

Figure 44: SCE Levels for Large and Small QSEs in Ramping Hours 123
Figure 45: SCE vs. Feasibility of Real-Time and Balancing Energy Schedules 124
Figure 46: Effects of Portfolio Ramp Constraints 128
Figure 47: Portfolio Ramp Constraints 130
Figure 48: Portfolio Ramp Constraints incorporating Schedule Changes 131
Figure 49: Actual Responsive Reserves Capability and Clearing Price 134
Figure 50: QSEs Failing to Satisfy their Responsive Reserve Obligations 136

LIST OF TABLES

Table 1: Zonal Shift Factors and Interzonal Redispatch Impacts 35
Table 2: Summary of Load and Generation Shift Factors 52
Table 3: Analysis of Redispatch Quantities 64
Table 4: Comparison of Costs from Market Solutions vs. Non-Market Solutions 74
Table 5: Effects of Local Deployments on Balancing Energy Prices 86
Table 6: Clearing Price, Load, Scheduled Energy and Balancing Energy 129
Table 7: Clearing Price, Load, Scheduled Energy and Balancing Energy Using Reformulated Ramp Constraint 133

Acknowledgements

We wish to acknowledge the input and numerous comments provided by the staff of the Market Oversight Division of the Public Utility Commission of Texas, including Parviz Adib, Eric Schubert, and Danielle Jaussaud. We are also grateful for the assistance of ERCOT in supplying the data used in this report and for the helpful comments provided by ERCOT staff, including John Adams and Young Li.

EXECUTIVE SUMMARY

This report provides an evaluation of ERCOT's market rules, operating procedures, and actions taken to maintain reliability and facilitate the competitive market. Accordingly, this report addresses four areas that pertain to market operations:

- ERCOT's Zonal Market and Interzonal Congestion;
- Local Congestion Management;
- Load Forecasting; and
- Real-Time Dispatch and Regulation Deployment.

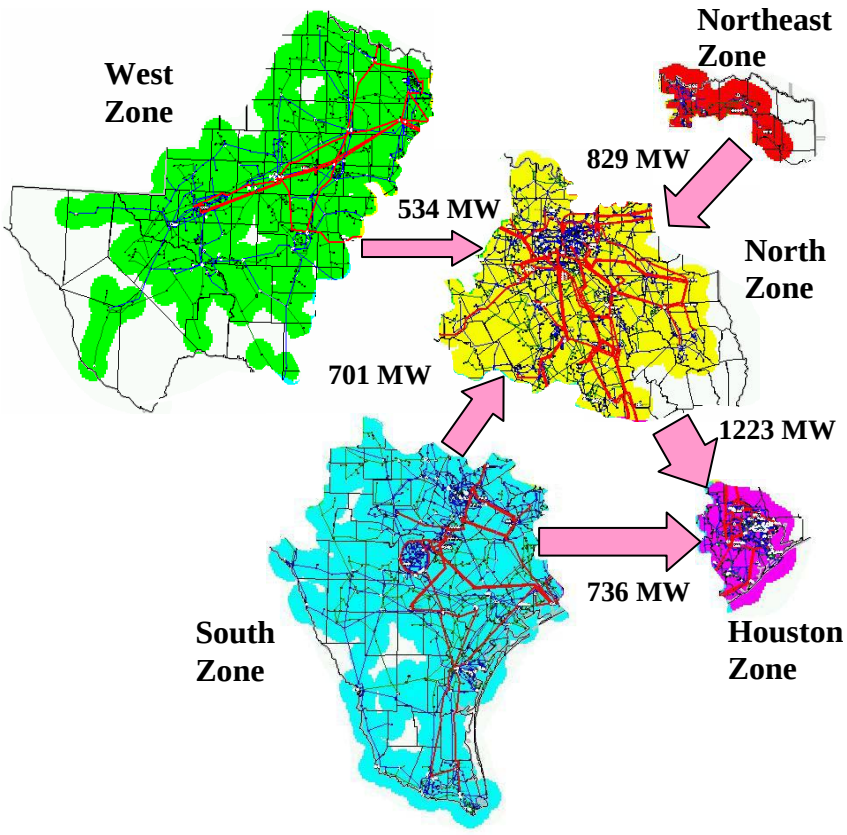
Based on the results of the analysis, we identify a number of areas of potential improvement and make recommendations to improve the performance of the markets. In the longer-term, the implementation of the Texas Nodal markets that are currently under consideration would more comprehensively address most of the issues identified in this report, particularly those related to managing congestion.

This executive summary will summarize the analytic results and findings of the report in each area. Our recommendations based on these findings are consolidated at the end of the executive summary.

A. Summary of Interzonal and Local Congestion Costs

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, congestion on the transmission network is managed in two ways. First, the ERCOT market is comprised of five zones that are interconnected by transmission interfaces referred to as Commercially Significant Constraints ("CSCs"), shown in the map below. The map also shows the physical transmission limits for each of the CSCs.

ERCOT Zones and CSCs -- 2004



The flows over the CSCs are managed by deploying balancing energy in each zone through the balancing energy market. The balancing energy market increases energy production in one zone and reduces it in another zone to reduce the flows between the two zones when a CSC is binding (i.e., when there is interzonal congestion). In any zonal market, zones should be defined that maximize the portion of the congestion that is managed by the zonal markets while still maintaining a manageable number of zones.

Second, constraints that are not defined as part of a CSC are referred to as local constraints and result in "local congestion" when they are binding. Local congestion is managed through the redispatch of individual generating or load resources. In particular, the actions that can be taken by ERCOT to manage local congestion include:

- Redispatching specific units out-of-merit order, referred to as out-of-merit energy ("OOME")[1];

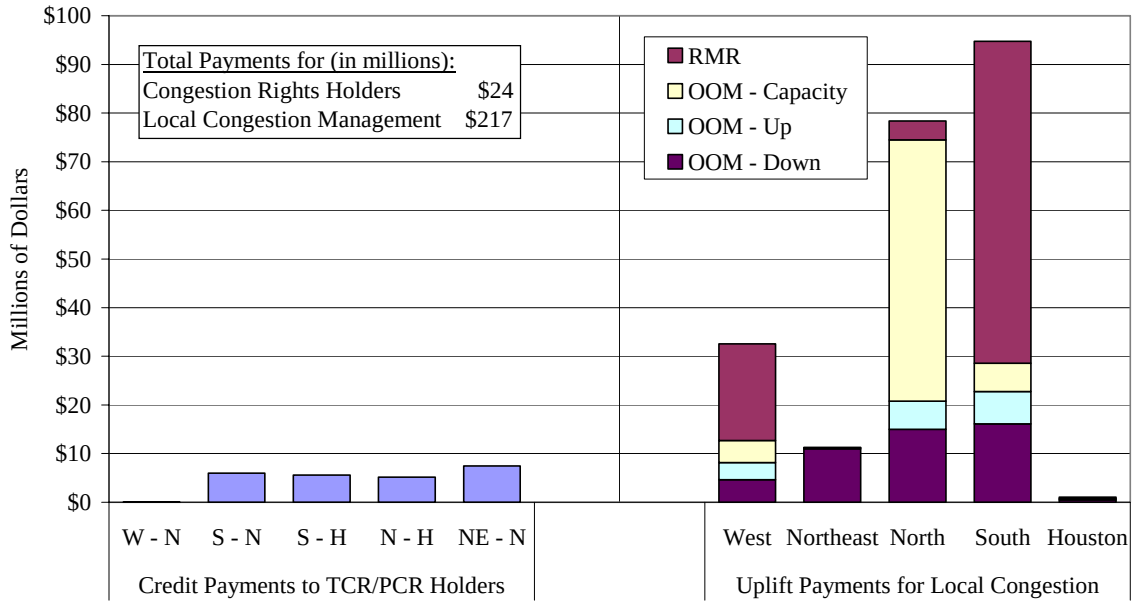
[1] Most OOME deployments occur by activating local constraints in ERCOT's balancing market model,

- Manually committing a resource out-of-market that will help relieve the local congestion, known as out-of-merit capacity ("OOMC"); and

- Committing or dispatching a reliability must-run ("RMR") resource. In general, units contracted to provide RMR service to ERCOT are compensated for start-up costs, energy costs, and are also paid a standby fee.

Local congestion costs arise from the payments made to suppliers that redispatch specific units to manage the congestion. These costs are collected through uplift charges to loads on an ERCOT-wide basis. Interzonal congestion costs occur when a CSC constraint binds in real-time, causing the zonal prices to differ to reflect the economic value of the constraint. A holder of a Transmission Congestion Right ("TCR") is entitled to receive the congestion value of the CSC constraint. Congestion rights payments are a useful way to measure the interzonal congestion costs. Hence, the figure below shows payments to CSC congestion rights holders compared to uplift payments to resources providing local congestion relief for the period January to August 2004.

Payments to TCR Holders vs. Local Congestion Payments
January to August 2004



generally referred to as "local balancing deployments". However, ERCOT operators can also issue manual dispatch instructions. Both means of redispatching individual resources result in comparable settlements with the resource owner and, hence, we refer to both types of local deployments as out-of-merit energy.

The left panel shows total payments to holders of congestion rights since the beginning of 2004. There have been virtually no payments to holders of congestion rights for the West-to-North CSC. However, payments for each of the other CSCs ranged from $5.0 million to $7.5 million through August 2004, totaling $24 million for all of the CSCs during the study period. The right panel shows uplift payments for each type of local congestion management in each zone. Despite reductions in local congestion costs from 2003, the local congestion costs of $217 million were almost ten times higher than the interzonal congestion costs. These costs are not directly comparable since local congestion costs are the incremental costs of redispatching individual units while the interzonal congestion costs reflect the market value of the congestion.

Nonetheless, the report includes a variety of analyses to compare the significance of local congestion and interzonal congestion in ERCOT, which all show that the majority of transmission congestion does not occur on CSCs. Local congestion that is not resolved using CSCs is not directly reflected in the zonal clearing prices. Hence, one may conclude that most of the economic value of congestion in ERCOT is not reflected in the zonal balancing energy prices. This raises significant concerns because it indicates that the current market prices are not efficiently and transparently revealing the costs of congestion in ERCOT. This has short-term effects on production and consumption, as well as long-term effects on investment and retirement decisions. The fact that most congestion costs are recovered through uplift charges that are socialized across the ERCOT region means that:

- Resources valuable for relieving network constraints will not receive compensation reflecting their value to the system, which will limit investment that would otherwise occur in congested areas;
- Resources that contribute to local congestion are effectively over-compensated because the zonal price they receive does not reflect the costs they impose;
- Loads that have an ability to respond to price signals will not receive accurate economic signals relating to their effects on local transmission constraints; and
- Loads and other market participants have a limited ability to hedge the costs of congestion they face.

Under the current market design, these issues can only be addressed by improving the definition of the zones and CSCs in ERCOT. In the longer-term, we believe that the

implementation of nodal electricity markets would most comprehensively resolve most of these issues because the prices at each location would accurately reflect both the local and interzonal constraints in the current system. In addition to improving the accuracy of the price signals, the nodal markets would allow the loads to fully hedge congestion costs by securing transmission rights and contracting bilaterally for delivery of power at specific locations. In the next section, we critically evaluate the interzonal congestion management processes and procedures.

B. ERCOT's Zonal Market and Interzonal Congestion

In any zonal market, there is a tradeoff between how much of the congestion is managed through the zonal market and the simplicity of the market framework. A small number of zones in the ERCOT market will result in fewer CSCs and a smaller share of the network congestion managed and priced through the balancing energy market. As the number of zones and CSCs increase, the share of the congestion costs reflected in the zonal balancing energy prices increases.

Some believe the increase in complexity associated with more zones and CSCs would reduce liquidity in the forward electricity markets. Hence, ERCOT currently uses a simplified network model with five zone-based locations and five CSCs. However, one new zone and two new CSCs were added as recently as 2004 to shift congestion costs from local to interzonal. The Northeast zone was created at the beginning of 2004 to allow the zonal market to capture local congestion occurring in the North zone. In addition, ERCOT created a new CSC from North-to-Houston to improve the management of congestion into Houston.

ERCOT's balancing energy market software, the Scheduling, Pricing, and Dispatch model ("SPD"), implements the zonal framework by making simplifying assumptions to represent the complex interactions among generators, loads, and the transmission network. Because the modeled power flows in a zonal market are based on simplifying assumptions, the estimated flows can depart significantly from real-time physical flows. Estimated flows that diverge significantly from actual flows can result in inaccurate congestion modeling and inefficient energy prices. Hence, if the system is operating

efficiently, the simplified modeling framework and assumptions should not cause large differences between modeled network congestion and the physical congestion on the network.

Zonal Congestion Management and Generation Shift Factors

When a resource produces output on an electricity network, the specific location of the resource and the load determines where the power actually flows. Resources in different locations have different impacts on the flows of power over various transmission facilities. The effect of a particular generator on a transmission facility (or a CSC) is referred to as a generation shift factor ("GSF"). Thus, a 20 percent shift factor for a particular resource implies that a 100 MW increase in output by the resource causes the flow over the interface to increase by 20 MW.[2]

In a zonal market, each resource within a zone is assumed to have the same shift factor relative to a CSC. This allows ERCOT to operate the balancing energy market with portfolio supply offers because the supply in each zone is assumed to be fungible. The zonal shift factors are computed by calculating the average GSF values of the generators in the zone. The zonal shift factors are re-calculated each month. The transmission cases used to calculate the monthly shift factors are also used to determine the limits for the CSCs and administer the sale of monthly transmission rights.

Using these zonal average shift factors, the balancing energy market model manages congestion over the CSC by increasing output in one zone and decreasing it in another zone. Resources located in zones that tend to positively impact flows are generally instructed to reduce output while ones in zones that negatively affect flows are instructed to increase output until the modeled flow is reduced to the flow limit.

[2] The total effect of a dispatch change on the flow over a transmission facility depends on both the source and the sink for the power. Hence, the incremental flow caused by one additional 1 MW that is produced at one location and consumed at another location would be equal to the sum of the shift factors for the generator (source) and the load (sink).

The SPD modeling of the system in the current balancing energy market can be inconsistent with the actual physical system for a number of reasons, including that:

- The individual generator shift factors will not match the zonal average shift factor so certain dispatch patterns within a zone can cause the model to diverge from reality;
- The distribution of the load in each zone does not match the generation distribution used to calculate the zonal shift factors, which will cause discrepancies in the CSC flows that change as the load increases and decreases; and
- The configuration of the system can change on an hourly basis (e.g., due to transmission outages) while the zonal shift factors are fixed for the month.

These inconsistencies are addressed in the real time by the ERCOT operators who adjust the CSC limits so that the CSCs will be binding in the model when the CSCs are physically binding in reality. This will cause the model to redispatch generation to limit the CSC flows, which will be effective even if inefficient.

The report includes a number of analyses to evaluate the performance of the current system. The first analysis focuses on the extent to which the shift factors for individual generators in each zone are consistent with the zonal average shift factor used by SPD. Based on this analysis, we find that the shift factors for the individual generators within some of the zones vary widely and, therefore, deviate substantially from the zonal average shift factors used to operate the balancing energy market.

For example, the generators in the North zone exhibited both positive and negative shift factors for each of the CSCs. In other words, incremental output from some generators in the North zone increases the flow on a given CSC while incremental output from other generators would relieve the flow on the same CSC. This is a significant departure from the assumption that all generators in the zone have the same effect on the flows over the CSCs.

Ultimately, the differences between the actual resource-specific shift factors and the zonal average shift factors result in inefficient congestion management and inaccurate zonal prices. If the model recognized the relative effectiveness of the resources in

affecting the flows over the CSC, it would redispatch different resources than are redispatched through the current market, which would generally result in lower, more efficient zonal prices. The dispersion of generation shift factors also raises concerns that suppliers may have the opportunities to create congestion on CSCs without incurring the associated costs. These conclusions and associated recommendations are discussed further below.

Beyond the analysis of the generator shift factors, the report includes a number of other analyses to evaluate the operation of ERCOT's zonal market. The results of these analyses lead to the following findings:

- There have been significant differences between the CSC interface limits used in the SPD model and the physical interface limits. In addition, the SPD limits for most of the CSCs have been highly volatile.
- Although the correlations between SPD flows and actual flows are generally strong, the differences are larger than we expected. The largest differences between modeled flows and physical flows on a monthly average basis ranged from 30 percent to 80 percent of the physical limits on the various CSCs. In a significant number of hours for three of the CSCs, the modeled flows traveled in the opposite direction of the physical flows.

Like the findings regarding the generation shift factors, these findings raise significant issues regarding how effectively the zonal market framework is able to reflect the reality of the transmission system and power flows in ERCOT. Although we provide some recommendations to improve the performance of the current markets, a market design that more accurately reflects the electricity system in ERCOT would ultimately be needed to fully address these issues and achieve the associated efficiency benefits.

Redispatch Analysis

The final analysis that we performed to evaluate the performance of ERCOT's zonal market is an analysis that evaluates the amount of generation needed to be redispatched to manage interzonal congestion.

The most important simplifications in ERCOT's zonal market are the use of portfolio supply offers and zonal shift factors to model the flows over the CSCs and manage

interzonal congestion. This is not efficient because the market is not able to select the most effective and economic resources within a zone to manage the congestion. When the resource-specific shift factors vary substantially, the optimal unit-specific redispatch to manage congestion can be much different than the zonal redispatch.

This report analyzes the significance of this simplification by evaluating how it affects the quantity of generation that must be redispatched to manage interzonal congestion. In particular, we perform a comparative analysis of the zonal market redispatch that actually occurred during the study period versus the economic redispatch that would have been possible on a resource-specific basis using a simulation model.

The results of this analysis show that:

- The quantity of generation redispatched under the current zonal market is double the quantities that would be redispatched if the most effective generators are redispatched, regardless of costs.
- The quantity of generation redispatched under the current zonal market is 30 to 60 percent higher than the quantities that would be redispatched economically utilizing resource-specific costs and shift factors.

These results likely understate the effects of moving to a full nodal electricity market as is currently under consideration in Texas. Under a nodal market, resource-specific characteristics would be considered in both the dispatch of generation and the commitment of the generation. Hence, units that may not be committed under the current market design due to their costs could be committed economically under a nodal market design because they are particularly effective at managing the flow on a constrained transmission facility.

Recommendations to improve the operation of the current zonal market are discussed in the final section of the executive summary together with the recommendations addressing other current ERCOT market issues.

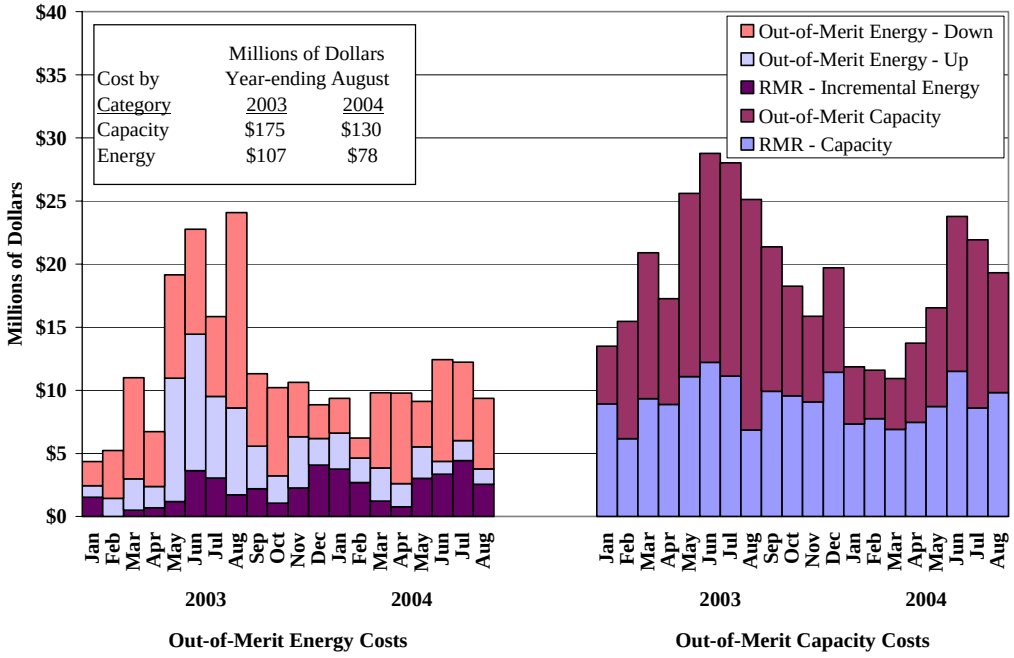
C. Local Congestion Management

As described above, ERCOT utilizes a number of different actions to manage congestion on transmission facilities that are not included in the CSCs. These actions provide

resource-specific instructions to suppliers, which include OOME Up or OOME Down dispatch instructions, OOMC commitment instructions, and RMR commitment and/or dispatch instructions. The costs of these actions are charged to all loads in ERCOT.

The following figure shows the costs for each of these actions by month for 2003 and the first eight months of 2004. The left side shows costs associated with dispatch instructions (OOME Up and OOME Down, and incremental energy from RMR units), while the right side shows the capacity costs associated with commitment instructions (RMR units and OOMC commitments).

Expenses for Out-of-Merit Commitment and Dispatch
2003 and 2004



The figure shows that out-of-merit energy costs declined by $29 million in the first eight months of 2004, a decrease of 27 percent. The reduction in OOME costs was driven primarily by a 61 percent reduction in OOME Up costs. The sum of out-of-merit capacity costs also decreased during the first eight months of 2004. The total costs for OOMC and RMR units decreased by $45 million, a decrease of 26 percent.

Out-of-merit costs are greater during the summer when higher loads increase the need for ERCOT operators to take out-of-merit actions to manage local congestion and reliability requirements. RMR costs did not increase substantially during the summer months

because RMR payments are primarily designed to recover fixed costs, which are constant throughout the year.

Although the costs are borne by load throughout ERCOT, the costs are incurred in specific locations on the network. The report includes a number of analyses showing where local congestion costs are incurred. These analyses show:

- The largest source of out-of-merit capacity costs is the Dallas-Ft. Worth area, although these costs decreased by more than $20 million during the first 8 months of 2004 versus the comparable period in 2003.
- The costs of out-of-merit dispatch during the study period in 2004 decreased by $48 million or 62 percent in the North, Northeast, and Houston zones compared to the same timeframe in 2003. A large share of the reduction in local congestion in these areas can be attributed to the creation of the Northeast zone in 2004.
- The uplift payments for out-of-merit energy in the Houston zone decreased 93 percent during the first eight months of 2004. This dramatic reduction in out-of-merit dispatch costs is related to the creation of the North-to-Houston CSC in 2004. The North-to-Houston CSC allows the balancing energy market to resolve the congestion on the 345 kV lines directly connecting the North zone to Houston.
- In the West zone and South zone, the costs of out-of-merit energy (including RMR costs) increased by $20 million during the first 8 months of 2004. More than one-half of this increase was paid to units under RMR contracts.

In summary, there have been significant reductions in expenses for out-of-merit commitment and dispatch actions in 2004. These reductions are due in part to the creation of a Northeast zone and the North-to-Houston CSC. These enhancements to the existing market shift more of the congestion costs from local congestion to interzonal congestion. This allows the zonal market framework to establish more transparent price signals and improves market participants' ability to hedge congestion costs.

However, most of the congestion costs in ERCOT are still associated with local congestion rather than interzonal congestion. A large share of the local congestion costs can be attributed to the RMR agreements. Such agreements are necessary, in part, because the energy prices in ERCOT do not reflect the economic value of local congestion. We believe that the Texas Nodal markets that are currently under consideration would provide a more robust pricing framework, allowing resources

currently under RMR agreements to primarily be compensated through a market that more fully reflects the value of their resources in relieving network constraints.

Multi-Step Congestion Management Process and the Balancing Energy Market

In addition to reviewing the patterns of local congestion and how those patterns have changed over the past two years, we evaluate ERCOT's local congestion management processes.

To resolve local congestion, ERCOT solves the balancing energy market in three steps. In the first step, the ERCOT software determines the dispatch levels based on zonal portfolio schedules and offers to meet demand while observing interzonal transmission limits. In the second step, incremental/decremental dispatch changes are made to specific resources to manage local congestion. In the third step, the software uses portfolio offers to counter-balance changes from the second step. This is done by re-clearing the balancing energy market while considering the interzonal and active local constraints from the second step. Our analysis in this report evaluates the ERCOT software and this multi-step process.

In an electricity network, all elements are inter-related. Even when done optimally, actions taken to manage local congestion can have substantial impacts on the balancing energy market outcomes. For example, when resources are dispatched down for local congestion, it reduces the supply of energy from those resources and causes additional balancing energy to be deployed. These effects can substantially increase prices under tight conditions in the balancing energy market.

To evaluate how large an effect local congestion had on prices in ERCOT during the study period, we analyzed 133 intervals from June through September 2004 when there was local congestion and the average balancing energy price was higher than $80/MWh. By re-running the market software without the local congestion in these intervals, we found that:

- In intervals when the price was higher than $150/MWh and units were turned down to manage local congestion, the average price increase caused by the local congestion was $123/MWh.

- In intervals when the original price ranged from $100/MWh to $150/MWh and units were turned down to manage local congestion, the average price increase caused by the local congestion was $62/MWh.

The current three-step congestion management process was developed presumably to preserve the integrity of the zonal market by separating the actions taken to manage local congestion from the actions taken to manage interzonal congestion. However, because the resource-specific deployments are counter-balanced with portfolio deployments, the true cost of decrementing or incrementing a unit for local congestion includes the cost of the associated increase or decrease in portfolio deployments. This is not considered in the current multi-step process, which can cause the model to make inefficient choices and cause artificial price spikes in the balancing energy market.

To address these issues, we recommend that ERCOT to modify the multi-step balancing energy market optimization to minimize the costs of deployments for both local congestion and interzonal congestion. The original logic underlying the multi-step process, that actions to relieve local congestion should not affect the balancing energy market, is fundamentally flawed. Hence, the current market should be improved by recognizing the significant interaction between local congestion management and the balancing energy market and making the necessary changes to SPD to make the entire process as efficient as possible. Doing this will significantly improve the efficiency of the pricing in the balancing energy market.

D. Load Forecasting

Load forecasting is important in two timeframes: day-ahead and real-time. The day-ahead forecast is used to determine when ERCOT must make supplemental commitments (through OOMC, replacement reserves, or RMR processes) to meet its reliability requirements. The day-ahead forecast is publicly available to market participants to assist them with commitment and scheduling decisions.

The real-time forecast, which is generally referred to as the short-term load forecast, is one of the primary inputs to the balancing energy market. The balancing energy market must make deployments to satisfy the short-term load forecast as adjusted by the offset entered by the ERCOT operators (known as "net load" or "SPD load").

Day-ahead forecasting

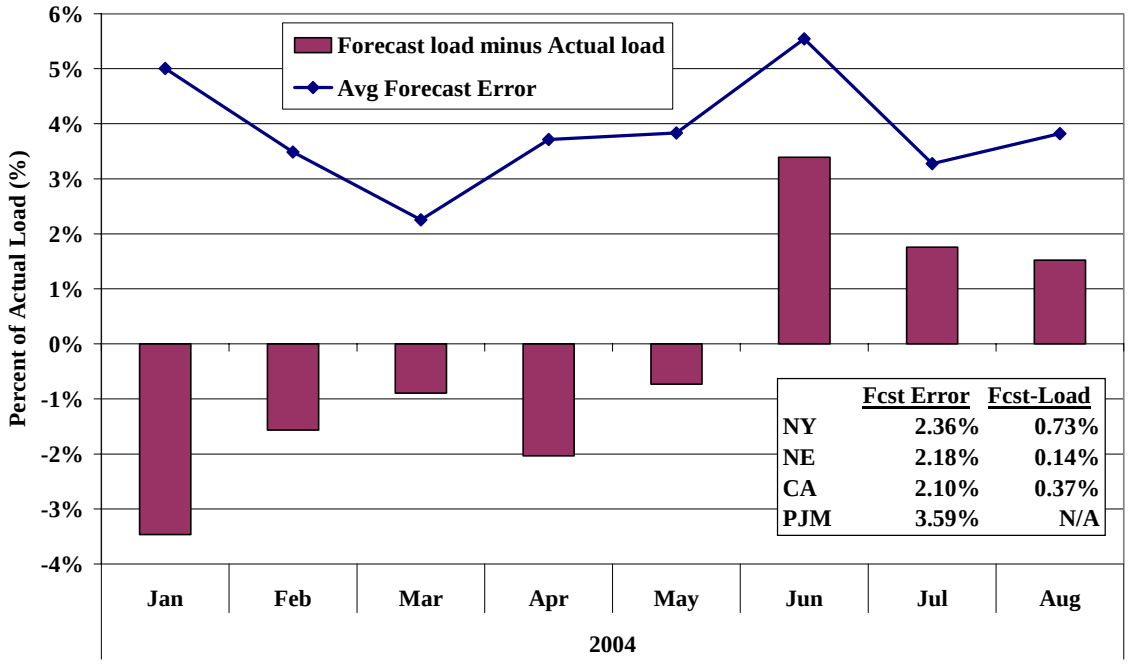
The analysis below includes two metrics to quantify the accuracy of forecasting:

- "Forecast-Actual": the average forecasted load minus the average actual load. This captures systematic differences between the forecast and the actual loads.
- "Forecast Error": the average of the absolute value of the difference between the forecasted value and the actual value ("forecast error"). This measures the magnitude of error regardless of the direction.

If all differences between the forecast load and actual load are random, the average forecast-actual metric can be close to zero while the forecast error is high.

To avoid emergency actions in real-time, the market must commit sufficient capacity to serve load and provide reserves under peak load conditions. Therefore, it is most important to accurately forecast load during the highest load hour. The following figure shows the accuracy of the day-ahead load forecast for January to August 2004 using a monthly average of the daily peak forecast hour.

Forecast Error at Daily Peak
January to August 2004



The bars in the figure show the forecast-load metric indicating that the forecast load was systematically lower in the winter months and in April, but higher in June. Over the

eight-month period, ERCOT over-forecasted load by an average of 0.11 percent. The inset in the figure presents comparisons to other ISOs. It shows that ERCOT's' over-forecasting is comparable to the ISO-NE but significantly lower than the New York ISO and CA-ISO.

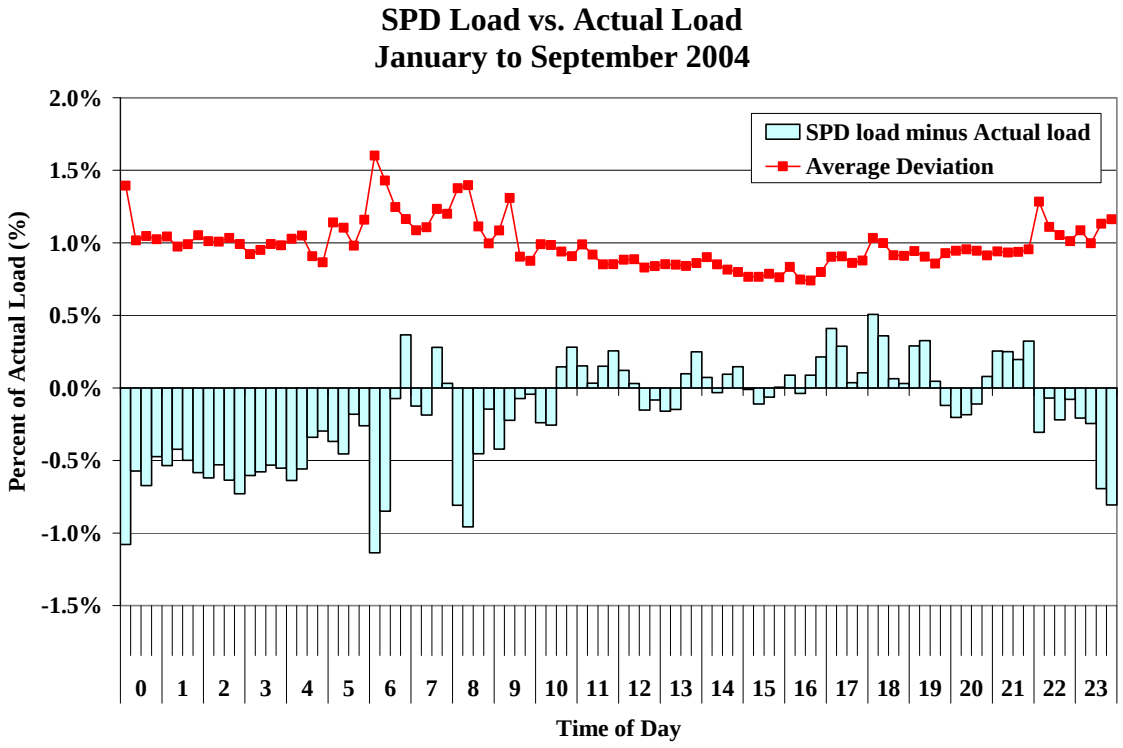
The figure also shows that ERCOT's forecast error for the eight-month period was 3.9 percent on average, which is higher than the average forecast errors for the other regions we examined. Based on these results, we recommend that ERCOT examine its load forecasting model to determine whether there are additional factors that could be considered or other changes that would improve its accuracy.

Real-time Forecasting

The balancing energy market model schedules energy to meet a load level specified by the operator every 15 minutes. The final load quantity, SPD load, is the sum of the short-term load forecast and operator offset. The short-term load forecasting model updates the forecast approximately 25 minutes prior to the start of each balancing interval. The operator offset is based on three factors:

- Updates to the load forecast made by the operator;
- Significant discrepancies between the QSEs' energy schedules and the planned generation in their resource plans; and
- Indications that the SPD load needs to be adjusted based on prior regulation deployments, system frequency, schedule control error ("SCE"), time-of-day factors, or other factors related to load.

Because a precise breakdown of the three parts of the offset was unavailable from ERCOT, we could not determine how the operator forecast differed from the short-term forecast. Hence, the report primarily focuses on the deviations between the SPD load and actual load. The following figure shows these deviations by time of day.



Like the day-ahead metrics, the figure shows two metrics that measure the differences between SPD load and the actual real-time load.

- The bars show the difference between the average SPD load and average actual load, which would reveal any systematic differences.

- The line shows the average deviation, calculated as the average of the absolute value of the deviation. We do not refer to this as a forecast error as we did for the day ahead because it includes the effects of the offset that are intended to address other factors. Hence, the differences are not necessarily errors.

Based on this analysis and others in the load forecasting section of this report, we find:

- The average SPD load is close to the average actual load during the afternoon, but systematically lower by 0.5 to 1.1 percent in the early morning and late evening.

- The differences between SPD load and actual load were largest near 6:00 am and 10:00 pm, the beginning and ending time for a large portion of bilateral energy contracts in ERCOT. The average deviations were also highest at these times.

- A significant portion of the deviations may be due to components of the offset that are intended to address factors other than discrepancies between the operator's forecast and the short-term load forecast, such as systematic schedule control error by suppliers.

These findings raise no significant issues and we, therefore, have no recommendations in the area of real-time forecasting. However, the finding that the differences are largest in the intervals close to 6:00 pm and 10:00 pm is consistent with a variety of operational issues that occur at these times. These operational issues are investigated and discussed in the next section.

E. Real-Time Market Operations

The fundamental requirement of the real-time operations is that supply continuously match demand. Failing to do so can result in blackouts, damage to electrical equipment, and other problems. To accomplish this, the real-time market and ERCOT operators take the following steps:

- Prior to each 15 minute interval:
  - The modeled load is determined (equal to the short-term load forecast plus the offset), which we refer to as SPD load.
  - The SPD model deploys the lowest cost balancing energy available to meet the SPD load.
- During the 15 minute interval:
  - Because the actual load will vary during the interval and generators may not produce the expected level of electricity (Schedule Control Error), ERCOT will deploy regulation on a 4 second basis to ensure that load matches generation.
  - However, regulation resources do not always respond reliably to the regulation signals and the signal itself can be inaccurate. This will create a residual error referred to as the Area Control Error ("ACE") that causes the frequency on the system to fluctuate.
  - If frequency fluctuates significantly enough, operating reserves will be deployed and under-frequency relays ("UFRs") will be tripped that curtail load to help restore the supply demand balance. These loads are referred to as Loads acting as Resources ("LaaRs").

This section of the report examines the efficiency of real-time operations and identifies areas for potential improvements.

Scheduling and Balancing Energy Market Outcomes

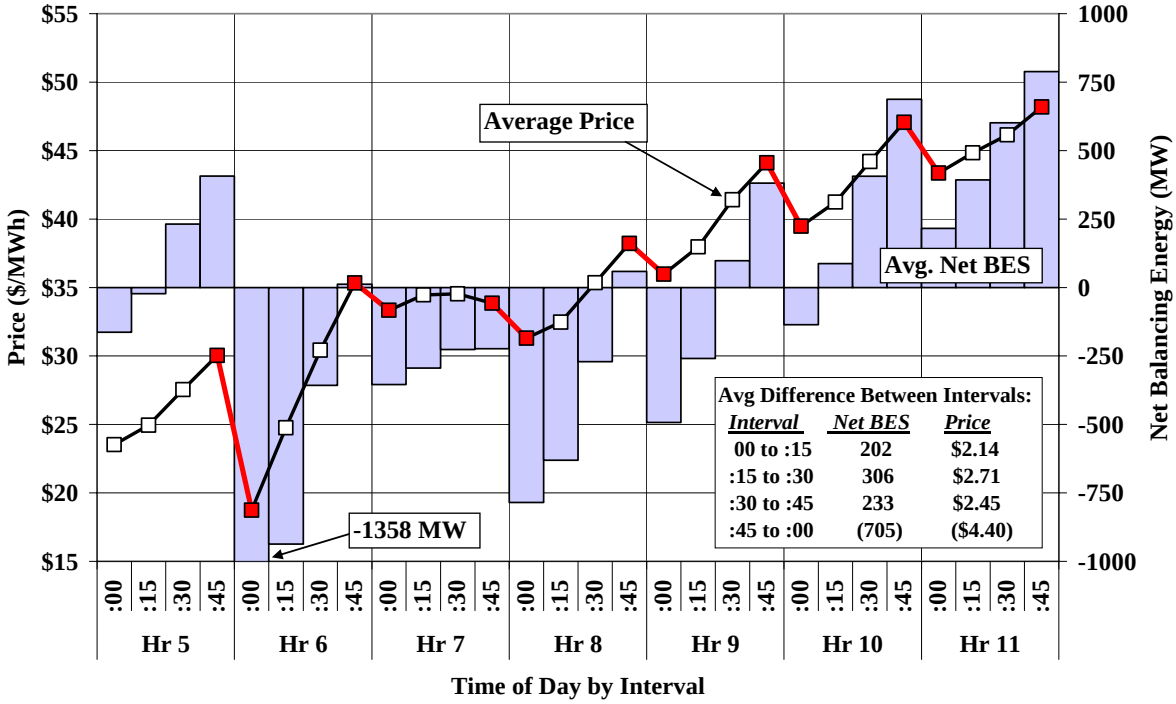
Our first analysis summarizes the performance of the balancing energy market, including an examination of scheduling patterns, balancing energy deployments, and balancing energy prices. This analysis is an update of a similar analysis that we presented in the 2003 State of the Market Report for ERCOT.

This analysis shows that although participants have the ability to submit energy schedules that vary every 15 minutes, only two of the largest QSEs tend to do so. The other QSEs generally submit schedules that change substantially at the top of each hour. In contrast, the load changes gradually over the hour. Because deviations between the load and the energy schedules are met by balancing energy deployments, these scheduling patterns can create extraordinary demands on the balancing energy market and erratic balancing energy prices. The following figures summarize these erratic price patterns by showing the balancing energy prices and balancing energy deployments in each 15-minute interval during the morning "ramping-up" hours and evening "ramping-down" hours.

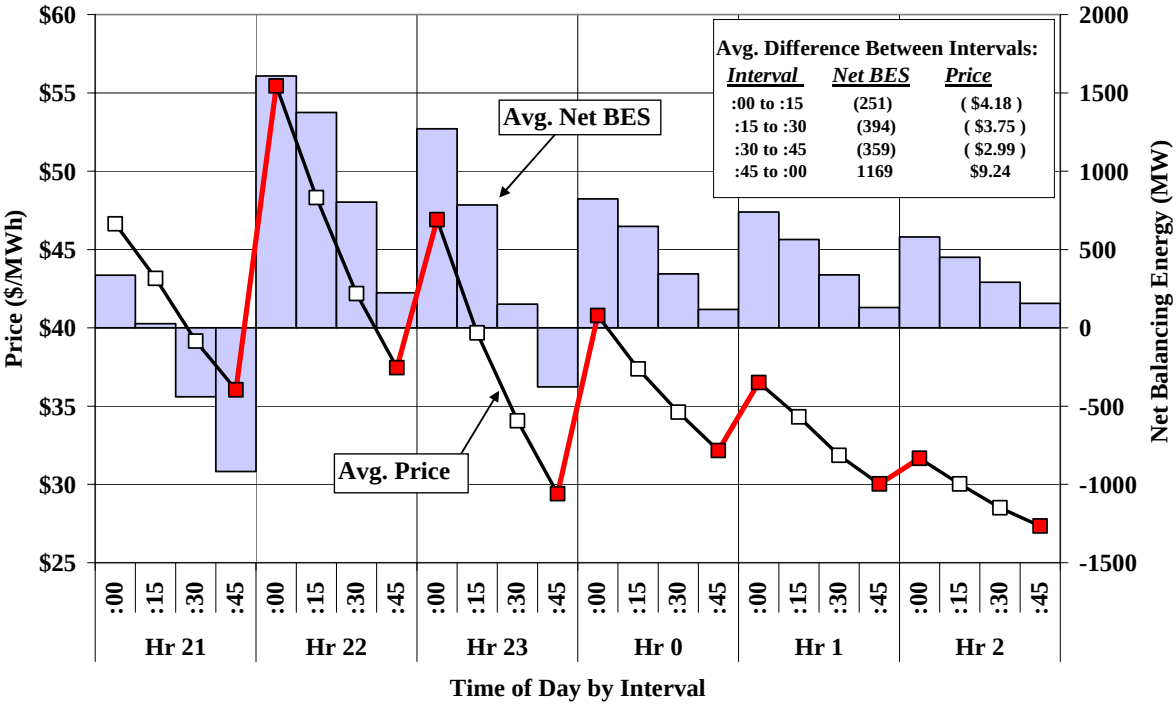
These figures show that the balancing energy deployments in the morning tend to fall sharply at the top of each hour before rising through the remainder of the hour. Balancing energy prices follow this pattern by falling sharply in the first interval of these hours and rising in the remaining three intervals. The opposite pattern occurs in the ramping down hours in the evening. In both the ramping up and ramping down hours, the largest jumps in deployments and prices occur at 6:00 am and 10:00 pm corresponding to the beginning and end of the peak hour period for bilateral contracts.

These balancing energy market results are primarily attributable to QSEs' inflexible energy scheduling patterns. The scheduling patterns that cause these balancing energy fluctuations also lead to other operational issues that we analyze in this report. As discussed in the final section of the executive summary, we make several recommendations to improve the QSEs' incentives and ability to schedule more flexibly.

Balancing Energy Prices and Volumes
Ramping-Up Hours – Jan. to Sept. 2004



Balancing Energy Prices and Volumes
Ramping-Down Hours – Jan. to Sept. 2004



Real-Time Operations and System Control

1. Regulation Needs in ERCOT

The balancing energy market clears every 15 minutes. However, the system must remain in balance continuously. ERCOT uses regulation resources to adjust output every four seconds in order to keep load and supply balanced.[3] The "regulation need" is the amount of regulation that would have to be deployed to keep supply and demand perfectly in balance. ERCOT carries as much as 1200 MW of regulation-up and regulation-down capability that it can deploy for this purpose.

The actual regulation deployment usually does not precisely equal the regulation need, either because the regulating units do not always accurately respond to the regulation signals, because ERCOT exhausts its regulation capability (i.e., the need is greater than the capability), or because the regulating units are limited in how quickly they can increase or decrease their output. The difference between the regulation deployment and the regulation need is the area control error. Positive ACE occurs when generators are over-producing and negative ACE indicates that generators are under-producing relative to actual load, including losses. ACE must be minimized to prevent equipment damage and maintain reliability. Our analysis of regulation need shows:

- Regulation need fluctuates significantly throughout the day, but is predictably more volatile during the morning load pick-up and evening load drop-off.
- Extreme quantities of regulation up are most frequently needed between 6:00 am and 6:30 am.
    - Almost two-thirds of the instances when more than 1400 MW of regulation up was needed occurred in this 30 minute period.
    - During the intervals-ending 6:15 am and 6:30 am, more than 1000 MW of regulation up was needed 5.5 percent and 6.7 of the time, respectively.

[3] In any electricity grid, differences between load and supply cause the frequency on the system to deviate from the ideal level of 60 Hz. Because ERCOT is separated from the eastern and western interconnects by DC ties, any over-generation or under-generation will translate directly to frequency deviations. This is not the case for control areas that are within the eastern or western interconnects where the generation in each control area will be continuously adjusted to maintain frequency throughout the interconnect. Hence, the deployment of regulation plays a particularly critical role in ERCOT for controlling the system and maintaining reliability.

- The largest quantities of regulation down need occurred both during the morning pick-up and evening load drop-off.
    - In the intervals-ending 6:00 am and 6:15 am, the regulation down need exceeded 1000 MW between 5.0 and 5.5 percent of the time.
    - During the evening load drop-off, more than 1000 MW of regulation down was needed 6 percent of the time in the interval ending 10:15 pm.

Currently, there is discussion whether to require larger amounts of regulation during the morning pick-up and evening drop-off in order to maintain frequency at these times. It is not a coincidence that the most frequent needs for very large quantities of regulation occur close to 6 am and 10 pm when the market participants in ERCOT are making the largest changes in their scheduled energy and balancing deployments are fluctuating widely.

Before deciding to increase the quantity of regulation, which will increase consumers' costs, it is important to understand the relationship of these factors and the extent to which they lead to problems controlling the system as evidenced by the ACE. The relationship of these factors and recommendations to improve system performance are provided in our analysis of the ACE summarized in the next subsection.

2. Area Control Error

While regulation deployments are generally effective at keeping supply and demand in balance, there are remaining differences because the deployments do not always match the regulation need. The extent to which supply and demand are out of balance is measured by the ACE. Our analysis of ACE shows that:

- ACE fluctuates significantly throughout the day. Like the regulation needs analysis in the prior subsection, the largest fluctuations occur in the morning load pick-up period and in the evening load drop-off period.
- There were a large number of negative ACE events between 6:00 am and 6:15 am. In more than 5 percent of the time in this interval, the ACE was lower than -450 MW (below the threshold that calls for ERCOT to deploy operating reserves).
- 41 percent of the days during the study period exhibited at least one instance of ACE lower than -450 MW and 38 percent of the days had one instance of ACE greater than 450 MW.

These results are troubling because it suggests that ERCOT frequently has difficulty controlling the frequency of the system for short periods of time. The two factors that contribute to the system ACE are: differences between generator obligations and actual output (schedule control error or SCE) and differences between SPD load and actual load (load deviations).

3. Schedule Control Error and Load Deviations

The report includes a number of analyses to quantify the SCE and load deviations in order to identify: a) the contributions of each to the regulation need and ACE in ERCOT, b) the underlying causes of these factors so that improvements can be made to reduce them, and c) QSE conduct that may be contributing to higher levels of SCE.

The results of these analyses showed that the SCE and load deviations tend to offset each other in general. This is consistent with expectations because the load deviations include the operator offset. One of the functions of the operator offset is to compensate for generators' SCE. This analysis showed that the load deviations tend to be negative before 5:00 am and after 10:00 pm, and positive throughout the middle of the day. This pattern serves to compensate for SCE levels that have to have the opposite tendency: to be positive before 5:00 am and after 10:00 pm, and negative throughout the middle of the day. The two largest QSEs exhibit SCEs that do not contribute to this pattern, exhibiting SCEs that are close to zero on average in most intervals.

However, the load deviations and SCE patterns are most critical in the intervals near 6:00 am and 10:00 pm because the regulation deployments and ACE are the most volatile during these intervals. This is due to the fact that load is changing rapidly during this period, as are the QSEs' schedules because many have bilateral contracts that span the peak period from 6:00 am to 10:00 pm. To examine the causes of the ACE and other system control issues that occur in the ramping periods, we focused a number of analyses of the load deviations and SCE levels on these intervals and found:

- SCE and load deviations are both much more volatile close to 6:00 am and 10:00 pm than at other times, contributing to the large fluctuations in regulation need and ACE at these times.

- The SCE fluctuations are much larger during the morning load pick-up period than during the evening load drop-off period.
    - Immediately before 6:00 am when large increases in energy schedules occur, generators accelerate their output and cause the average SCE to be relatively large and positive.
    - Shortly after 6:00 am, the average SCE becomes substantially negative because suppliers do not increase output fast enough to satisfy their schedule changes.
    - Some QSEs do not have the physical capability to increase their output fast enough to meet their energy schedule changes at 6:00 am and 10:00 pm.
- During the morning load pick-up and evening load drop-off, the SCE of other QSEs tend to peak at the top of the hour before decreasing sharply. Generally, this pattern results from the fact that all but the largest two QSEs change their schedules at the top of every hour.
- Although the SCE fluctuations are significant, the load deviations tend to be the larger contributor to the high regulation need and ACE during the morning load pick-up and evening load drop-off periods.
- The primary cause of the load deviations is how the SPD load is modeled over the interval.
    - Load and Generation are assumed to be ramping in the first and third 5-minute portions of the 15-minute interval, and flat in the middle 5-minute period.
    - In the middle 5 minutes of each interval during the morning hours, the load deviation decreases sharply. This occurs because the SPD load is assumed to be flat while the actual load is increasing rapidly.
    - Similarly, in the middle of each 15-minute interval during the evening hours, the load deviation increases sharply.
    - Because operators submit load to SPD every 15 minutes, they are limited in their ability to reduce the load deviations.
- Although SCE and load deviations tend to offset one another in most intervals, they tend to move in the same direction around 6:00 am, jointly contributing to the relatively large ACE and the need for regulation.

We make several recommendations at the end of this section that should reduce the load deviations and SCE levels.

Portfolio Ramp Constraints in SPD

In electric power markets, there are physical limits to the rate at which each resource can increase output. There are times when demand increases so rapidly that it is not possible for the least expensive unloaded resources to respond in economic merit order. Because resources are offered in the balancing energy market in portfolios, each QSE must submit a portfolio ramp constraint that allows the QSE to manage how quickly the balancing energy market can request increases or decreases in output from the QSE's portfolio.

It is important to accurately represent ramp limitations in order for the market to fully utilize the supply. When there are large changes in balancing energy deployments, which frequently occur in the morning and evening ramping periods, the QSE's ramp constraints can cause a large quantity of energy to be unavailable to the market and result in sharp balancing energy price changes. As a result, ramp constraints are a primary cause of price spikes that have occurred in the balancing energy market. During the study period, prices increased in 13 instances by more than $100/MWh from interval-ending 10:00 pm to interval-ending 10:15 pm.

We examine the current method of reflecting ramp constraints in the balancing energy market model. Our analysis reveals that ERCOT's current methodology limits how quickly deployments of balancing energy can change from interval to interval based on the QSEs' portfolio ramp constraint, ignoring the QSEs' energy schedule changes. This is significant, particularly during the morning and evening ramp periods when energy schedules are changing by large quantities each hour.

Balancing energy deployments are made incrementally to a QSE's energy schedule. Hence, a QSE whose energy schedule is increasing sharply will have a much greater ability to balance down than ERCOT's model will recognize and less ability to balance up. QSEs can protect themselves against the second issue (that they have less ability to balance up when their energy schedule is increasing) by reducing their ramp rates or offering less energy in the balancing energy market. This effectively reduces the supply of energy in the balancing energy market.

Similarly, a QSE whose energy schedule is decreasing sharply will have a much greater ability to balance up than ERCOT's model will recognize and less ability to balance down. The issues created by this modeling approach can be remedied by considering the QSEs' energy schedule changes when applying the ramp constraints in the balancing energy model. Hence, we recommend that ERCOT consider a specific modification of its portfolio ramp rate methodology.

QSE Provision of Reserves

In order to maintain the reliability of the ERCOT system, ERCOT requires 2300 MW of responsive reserves (i.e., 10-minute synchronous reserves). ERCOT satisfies this requirement through resources that are self-scheduled by QSEs and by procuring responsive reserves through the responsive reserve auction market conducted one day prior to the operating day.

QSEs' responsive reserve schedules must be satisfied by setting aside sufficient capacity to respond to a reserve deployment. We evaluate whether the system requirements are being satisfied, as well as whether individual QSEs are meeting their responsive reserve obligation. The results of our analyses indicate:

- The system generally holds 1500 MW to 3000 MW of responsive reserves in excess of the 2300 MW required.
- We found six QSEs that were short of their responsive reserves obligations a significant portion of the time. However, the quantities not provided were less than one percent of the total responsive reserve requirements on average.
    - Some of these shortages disappear if the 20 percent limit on the quantity of reserves that can be supplied by a single unit is relaxed.
    - We believe the 20 percent limitation is overly restrictive for small units.

Based on these findings, we have identified no significant concerns regarding the supply of responsive reserves in ERCOT, although this analysis does not evaluate whether QSEs respond reliably to responsive reserve deployment instructions from ERCOT. Nevertheless, we recommend that ERCOT modify the supply restriction to make it less constraining to relatively small units and institute procedures to monitor whether QSEs are meeting their reserve obligations in real time.

F. Summary of Conclusions and Recommendations

This section consolidates the recommendations we make in each of the areas we investigate.

Congestion Management

Our analysis of ERCOT's zonal market and interzonal congestion management process indicates that:

- Interzonal congestion has been relatively infrequent and the costs have been modest;
- The zonal market framework used to manage interzonal congestion makes important simplifying assumptions; and
- These simplifying assumptions can significantly affect the outcomes of the markets and can result in inefficient dispatch of the generating resources.

Unfortunately, many of these issues are inherent in the zonal market approach so potential improvements are relatively limited. Our primary recommendation for addressing the issues and concerns identified in this report is that ERCOT adopt the nodal electricity markets that are currently under consideration. Well-structured nodal markets would resolve most of the operational and efficiency issues that plague the current markets due to the zonal simplifications, the portfolio scheduling and bidding framework, and local congestion management procedures.

Absent implementation of nodal markets, we recommend the following changes to the current markets to improve the management of interzonal congestion:

1. Improve the process for designating zones to minimize the effects of the simplifying zonal assumptions.
2. Improve the process for evaluating and revising CSC definitions.
3. Modify the calculation methodology of the zonal average shift factor to exclude generation whose output is generally fixed (e.g., nuclear units).
4. Provide ERCOT the operational flexibility to temporarily modify the definition of a CSC associated with topology changes.

In the area of local congestion management, we concluded that local congestion management can have large indirect effects on portfolio energy deployments and the balancing energy prices. The current multi-step process does not efficiently consider the interaction between actions taken to resolve local congestion versus those taken to resolve interzonal congestion, resulting in inefficient market results and artificial price spikes in the balancing energy market. To address this concern, we recommend:

5. ERCOT modify its multi-step balancing energy market optimization to recognize the interactions between its local congestion management and zonal balancing energy deployments to minimize the costs of both classes of deployments.

Load Forecasting

The only significant issue we identify in our evaluation of ERCOT's load forecasting is that the accuracy of the day-ahead forecast was less than that of system operators in other regions and that the forecast errors seemed to exhibit a specific pattern over the day. Based on these findings we recommend:

6. ERCOT examine its load forecasting model to determine whether there are additional factors that could be considered or other changes to their model that would improve its accuracy.

Balancing Energy Market Operations

Consistent with the 2003 State of the Market report, we found that participants' hourly scheduling patterns contribute to erratic deployments and prices in the balancing energy market, particularly in the morning load pick-up and evening load drop-off periods. To address this issue, we recommend ERCOT consider changes that may increase the willingness of QSEs to submit flexible schedules (i.e., schedules that can change every 15 minutes). To that end, we have recommended that ERCOT consider introducing two scheduling options for participants:

7. Allow QSEs to submit an energy schedule for the end of the next hour that would be used by ERCOT to produce 15-minute schedule quantities by interpolating across the hour.

8. Implement an optional capability for QSEs to automatically adjust their hourly balancing energy offers for the changes in their 15-minute schedules. This

would help ensure that the participants' portfolio energy offers are consistent with their energy schedules when the energy schedules are changing each interval.

These changes would likely increase the portion of the load that is scheduled flexibly and improve the performance of the balancing energy market.

We identified two factors that contribute to the large fluctuations in the regulation need and ACE: SCE and load deviations. Although both factors are significant, the load deviations are a larger contributor to the system control issues. To reduce the magnitude of the load deviations, we recommend the following changes:

9. Eliminate the load and generation plateau in the middle of the interval by changing the modeling approach for load and generation.

This change would have the added benefit of making more capability available to the balancing energy market by lengthening the ramping periods from 10 minutes to 15 minutes.

With regard to SCE, we have two recommendations for ERCOT to consider that should reduce SCE levels:

10. Require that QSEs submit physically feasible energy schedules. It would be difficult to automatically validate the schedules for feasibility since they can be submitted well before the operating hour. However, it could be monitored and enforced through ex post validation of compliance.

11. Implement uninstructed deviation charges that allocate a portion of the regulation costs to the QSEs exhibiting large SCEs in the periods during each hour with the largest regulation needs. In particular, we propose specific changes to PRR 356 that have been pending for more than two years.

The final two operational issues evaluated in this report relate to the treatment of portfolio ramp rate limitations in the balancing energy market and QSEs' provision of responsive reserves. Regarding the ramp rate limitations, the report evaluates ERCOT's current methodology of applying the ramp rate limitations to balancing energy deployments without considering changes to QSEs' energy schedules. We conclude that this methodology can lead to significant distortions in the balancing energy market in

intervals with relatively large schedule changes and to significant changes in balancing energy deployments. To address this issue, we recommend the following:

12. ERCOT stakeholders consider a specific modification of ERCOT's portfolio ramp rate methodology that would recognize the changes in energy schedules.

With regard to responsive reserves, we identify isolated instances when individual QSEs are short of their responsive reserve requirements. Based on this analysis, we have no significant concerns regarding the supply of responsive reserves in ERCOT, although we did not evaluate QSEs' responses to responsive reserve deployments. Although we found no significant concerns during the study period, we recommend the following changes:

13. Modify the limit on the quantity of responsive reserves provided from a single unit to make it less binding on relatively small units.

14. Institute procedures to monitor whether QSEs are meeting their reserve obligations in real time.

I. INTERZONAL CONGESTION MANAGEMENT

A. Background

One of the most important functions of any electricity market is to manage the flow of power over the transmission network by limiting additional power flows over transmission facilities when the flows reach the facilities' operating limits or become unstable. In ERCOT, congestion on the transmission network is managed in two ways. First, the ERCOT market is currently comprised of five zones that are interconnected by transmission interfaces referred to as Commercially Significant Constraints ("CSCs"). ERCOT manages flows over the CSC interfaces in real time by deploying balancing energy in each zone through the balancing energy market. The balancing energy market increases energy production in one zone and reduces it in another zone to reduce the flows between the two zones when a CSC is binding (i.e., when there is interzonal congestion).[4]

Second, constraints within each zone that are not defined as part of a CSC are referred to as local constraints and result in "local congestion" when they are binding. ERCOT manages local congestion through the redispatch of individual generating resources. In this section of the report, we evaluate patterns of interzonal congestion and analyze the efficiency of ERCOT's congestion management processes and procedures.

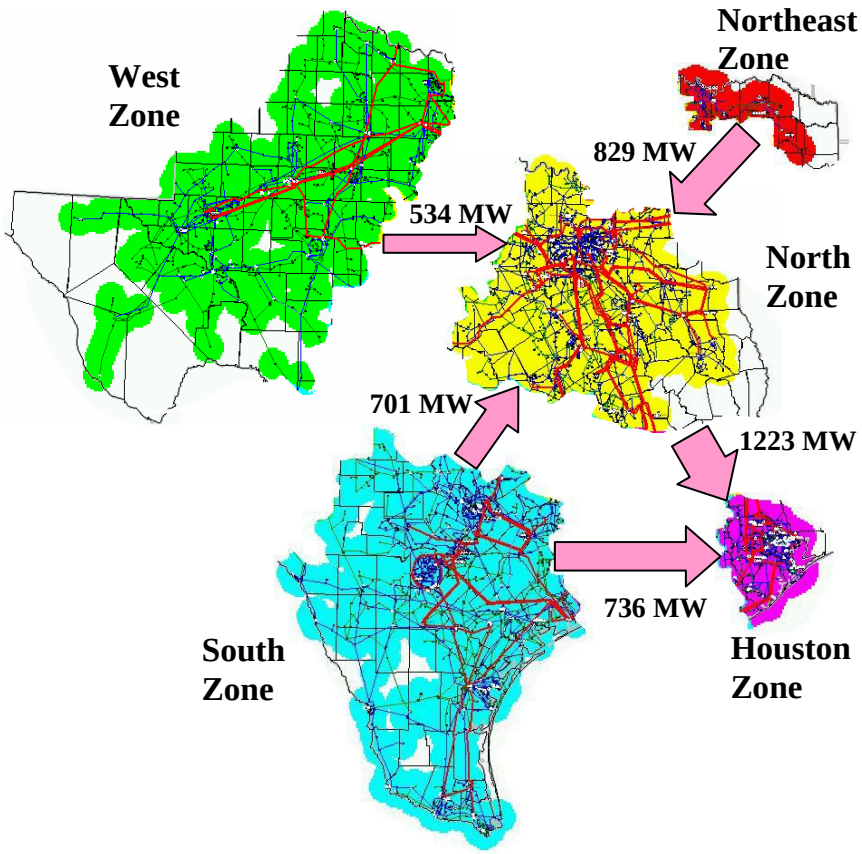
When approving this type of zonal market in 2001, the Public Utility Commission of Texas ("the PUCT") accepted a tradeoff where ERCOT would charge congestion rents on a limited number of transmission interfaces in order to provide market participants with a simpler market framework than the nodal market framework operating in other regions. The congestion that occurs over the CSCs is much more transparent than local congestion because it is revealed through zonal price differences that reflect the value of the interzonal congestion. Alternatively, ERCOT pays generators directly when it redispatches to manage local congestion, the costs of which ERCOT recovers through uplift charges to QSEs on a load-ratio share basis.

[4] These deployments are incremental to those necessary to balance supply and demand.

A smaller number of zones will result in fewer CSCs and a smaller share of the congestion being managed and priced in the balancing energy market. With a larger number of zones, the share of transmission congestion that ERCOT manages and prices through the balancing energy market increases and the amount of congestion that ERCOT must manage locally decreases.

However, increasing the number of zones can increase the number of CSCs geometrically, resulting in a more complex market with transmission rights distributed over a much larger number of transmission interfaces. ERCOT currently uses a simplified network model with five zone-based locations and five CSCs. The CSCs represent groups of transmission elements. ERCOT operators use planning studies and real-time information to set limits for each CSC that are intended to approximate the total transfer capability of the CSC. Figure 1 shows a map of the five zones and the five CSCs in ERCOT. The figure also shows the physical limits for each of the CSCs.

Figure 1: ERCOT Zones and CSCs



The Northeast zone was created at the beginning of 2004 to cause the zonal market to capture local congestion within the North zone. This local congestion increased considerably in 2003 when a new generator went online in that part of ERCOT, creating a chronic export constraint (or generator pocket) from what is now the Northeast zone into the North zone. The new Northeast-to-North CSC allows ERCOT to price the congestion and allocate the transmission capacity among the generators behind the constraint on an economic basis. In addition, ERCOT created a new CSC from North-to-Houston to improve the management of congestion into Houston.

ERCOT's balancing energy market model, the Scheduling, Pricing, and Dispatch model ("SPD"), implements the zonal framework by using simplifying assumptions specified in the ERCOT protocols that are intended represent the complex interactions between generators, loads, and the transmission network. Because the model's power flows are based on zonal approximations, the SPD estimated flows can depart significantly from real-time physical flows. Estimated flows that diverge significantly from actual flows can result in inaccurate congestion modeling that leads to inefficient energy prices, inefficient dispatch of resources, and other market costs.

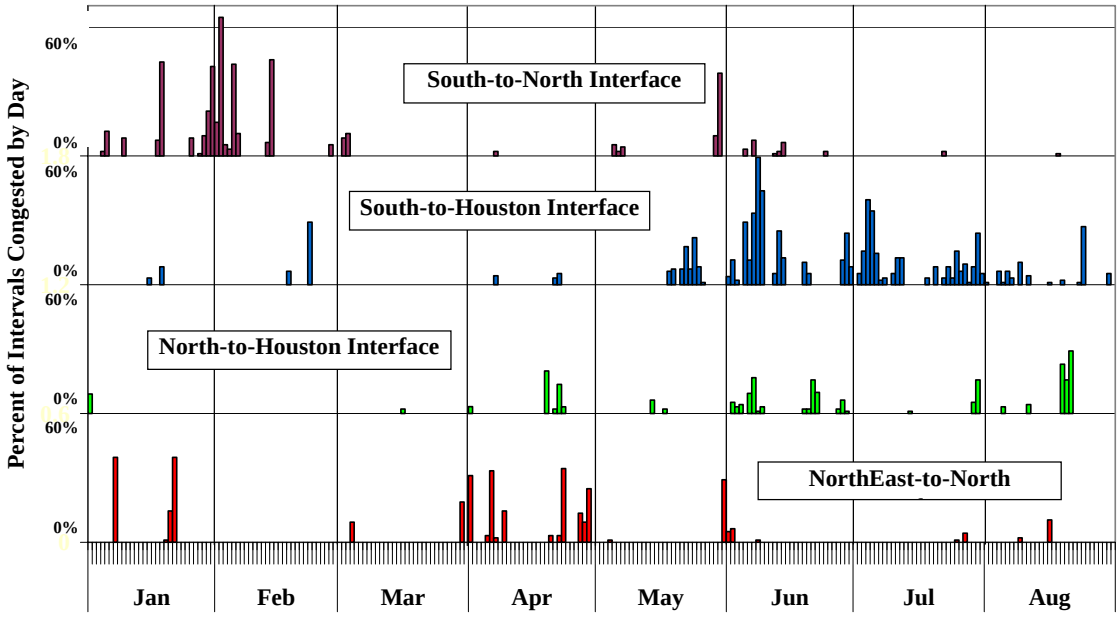
If the zonal market and other congestion management processes are operating efficiently, the simplified modeling framework and assumptions should not cause large differences between modeled congestion and the physical congestion on the network. In addition, managing most transmission congestion using CSCs is desirable since the zonal prices will reflect the economic value of the congestion and ERCOT directly assigns the congestion costs. Ideally, the deployment of units for local congestion should not significantly impact the supply-demand balance or zonal market outcomes in the balancing energy market.

B. Summary of Interzonal Congestion

In this subsection, we first summarize the patterns of interzonal congestion that occurred during the study period, then evaluate the operation of the zonal market and the efficiency with which it manages the flows between zones. To provide a preliminary view of the patterns of congestion that occurred on each of ERCOT's CSCs during the

study period, Figure 2 shows the frequency of congestion by CSC on a daily basis from January to August 2004. This figure does not show the West-to-North CSC because the CSC was rarely congested during the study period.[5]

Figure 2: Daily Percentage of Intervals Congested by CSC
January to August 2004



For each of the CSCs, congestion occurred no more than 3.1 percent of intervals. Excluding the South-to-Houston interface, congestion occurred less than 2.0 percent of the time. The most congested CSC, South-to-Houston, exhibited congestion most often during early June and July. The South-to-North interface was congested most frequently in late January and early February. Congestion on the North-to-Houston and Northeast-to-North interfaces occurred infrequently over the study period.

Based on the patterns shown in Figure 2, interzonal congestion has tended to occur sporadically on each of the CSCs. In fact, the most frequently congested CSC (South-to-Houston) experienced less than two intervals of congestion on 76 percent of the days during the study period. The South-to-North, North-to-Houston, and Northeast-to-North interfaces experienced less than two intervals of congestion on 87, 88, and 91 percent of

[5] There were only four intervals out of more than 20,000 with congestion on the West-to-North interface.

the days during the study period, respectively. In addition, the figure does not reveal any consistent seasonal patterns of interzonal congestion.

The sporadic pattern in congestion suggests that occurrences of congestion are frequently related to irregular system conditions such as transmission outages and generation outages. In most hours, a large share of the transfer capability of each CSC is already being used, so factors such as outages that affect the physical limits moderate amounts can have a large impact on the frequency of congestion. For instance, most of the congested intervals on the South-to-Houston CSC that occurred in late June coincided with the Brehnam North-Highway 36 138kV, the Waller-Chapel Hill 138 kV, and the Gliden-Bellville South 69 kV lines being out of service at various times. These outages resulted in transfer capability reductions over the South-to-Houston CSC averaging 235 MW. Likewise, a large share of the congestion on the Northeast-to-North CSC occurred on a handful of days where the physical transfer capability of the interface was reduced. Almost one-half of the congested intervals occurred on just seven days where the physical capability was reduced by 25 to 60 percent. Thus, factors that cause relatively small reductions in capability or increases in flows play a significant role in the occurrence of congestion.

C. Zonal Congestion Management and Generation Shift Factors

The balancing energy market model manages congestion over the CSC by redispatching resources on one or both sides of a constraint. ERCOT instructs resources that cause the most flow across a congested line to reduce output while instructing ones that cause the least flow (or provide counter-flows) across a congested line to increase output until the flow is reduced to the flow limit. The effect of a particular generator on a transmission facility (or a CSC) is referred to as a generation shift factor ("GSF").

For example, a 20 percent shift factor for a particular resource implies that if 100 MW is produced by the resource, the flow over the interface increases by 20 MW. Similarly, a -15 percent shift factor for a particular resource implies that if 100 MW is produced by the resource, the flow over the interface decreases by 15 MW. Since power must have a

source and a sink to flow, the GSFs for each generator are calculated assuming the generator is the source and assuming one or more reference points as the sink.

In a zonal market, each resource within a zone is assumed to have the same shift factor relative to a CSC (i.e., the same impact on the flow across the CSC per MW increase in output from the resources). This assumption allows ERCOT to operate the balancing energy market with portfolio supply offers because the supply in each zone is assumed to be fungible. The zonal shift factors are computed by calculating the average GSF values of the generators in the zone. Each month, ERCOT uses transmission planning studies to determine zonal shift factors for the following month. The same transmission cases are used to determine the limits for the CSCs and administer the sale of monthly transmission rights. The updated monthly shift factors adjust for known changes in the topology of the system due to planned transmission outages and upgrades.

To show how the generation shift factors for each zone together determine the assumed effect of an interzonal redispatch, we calculated the shift factors for the "source zone" and the "sink zone" for each CSC. For example, the source zone for the North-to-Houston CSC is the North zone and the sink zone is the Houston zone. Table 1 shows these calculations. The source zone shift factors tend to be the largest positive shift factors for the CSC and the sink zone shift factors tend to be the lowest shift factors.

Table 1: Zonal Shift Factors and Interzonal Redispatch Impacts

Interface	Source Zone Impact (1)	Sink Zone Impact (2)	Total Impact = (1) - (2)
West-to-North	40.3%	1.0%	39.4%
South-to-North	39.1%	0.6%	38.4%
South-to-Houston	18.4%	-17.9%	36.4%
North-to-Houston	0.4%	-34.9%	35.3%
NorthEast-to-North	35.4%	-5.0%	40.4%

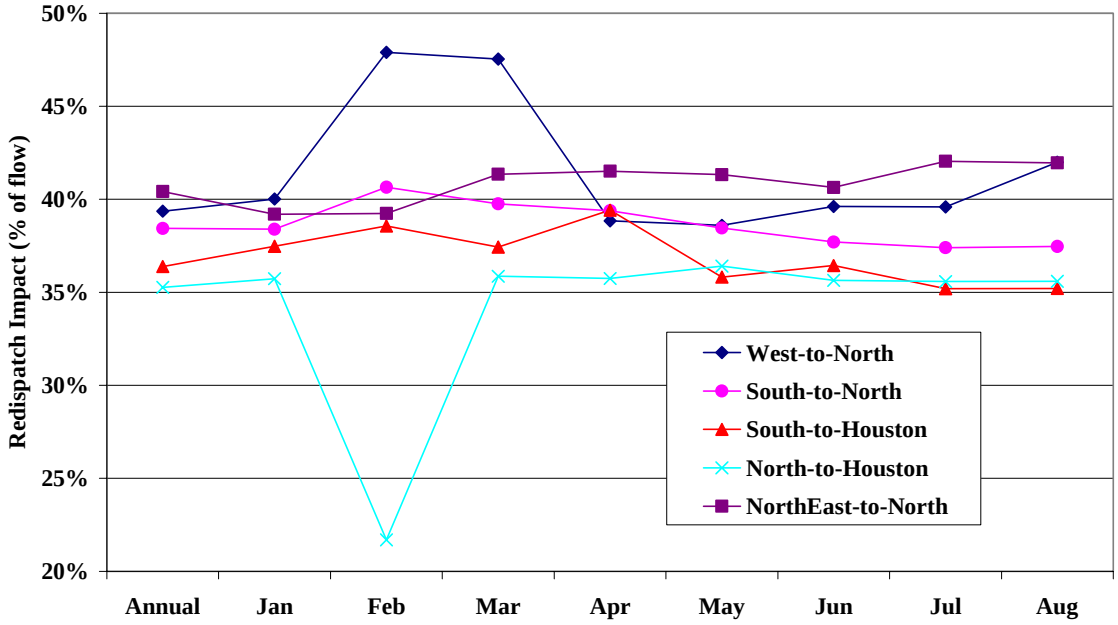
The results shown in the last column indicate a total impact of zonal redispatch ranging from 35 to 40 percent. For instance, the first row indicates that decreasing 10 MW in the West Zone and increasing 10 MW in the North Zone would reduce the flow over the West-to-North interface by 3.9 MW. In this case, the shift factor for the sink zone is actually positive, implying that additional generation in the North Zone increases flows across the interface. This is explained by the fact that a limited number of resources in the North create large positive loop flows over the CSC, although most of the resources in the North Zone actually reduce the flows on this interface (i.e., have a negative GSF).

The zonal shift factors used in the SPD model determine which portfolios ERCOT should redispatch to resolve interzonal congestion. As described above, the zonal shift factors are re-calculated each month to account for changes to the topology of the transmission grid. Figure 3 shows how the total redispatch impact implied by the zonal shift factors have changed from month to month during 2004.

Figure 3 indicates that the total redispatch impacts have remained relatively stable over the year, although there were isolated fluctuations early in the year. The annual estimate for the West-to-North interface redispatch impact was 39 percent, but rose to approximately 47 percent during February and March. Likewise, the redispatch impact for the North-to-Houston decreased to 22 percent in February from annual average of close to 35 percent. These changes are most likely due to planned transmission outages that occurred in these months.

Because the zonal shift factors change monthly while physical conditions on the network can change hourly, zonal shift factors will be inconsistent with prevailing electrical conditions on the network in some periods. This can affect the zonal redispatch and pricing. One of the primary means for addressing this inconsistency is to adjust the CSC transfer limit in SPD to ensure that the CSC will be binding in the model when the CSC is physically binding in reality. This will cause the model to redispatch generation to limit the CSC flows, which will be effective even if it is not optimal.

Figure 3: Total Redispatch Impact by CSC
January to August 2004



For example, if the physical limit for a CSC is 1,000 MW and the flow is 1,000 MW, the CSC is binding physically. If the SPD-calculated flow in this case was 600 MW, the ERCOT operator would have to set the limit for the CSC in the SPD model at 600 MW to cause the constraint to be binding and generation to be redispatched by SPD to manage the flows over the CSC consistent with the physical realities of the network.

Prior to the beginning of each month, ERCOT uses planning studies to update zonal-average shift factors and the estimated physical limit of each CSC. The planning model's estimate of the real-time transfer limit of each CSC is referred to as an operating constraint 0 limit ("OC 0 limit") while the transfer limit that ERCOT actually uses in real-time is known as an operating constraint 1 limit ("OC 1 limit"). The OC 1 limit will differ from the OC 0 limit for two reasons.

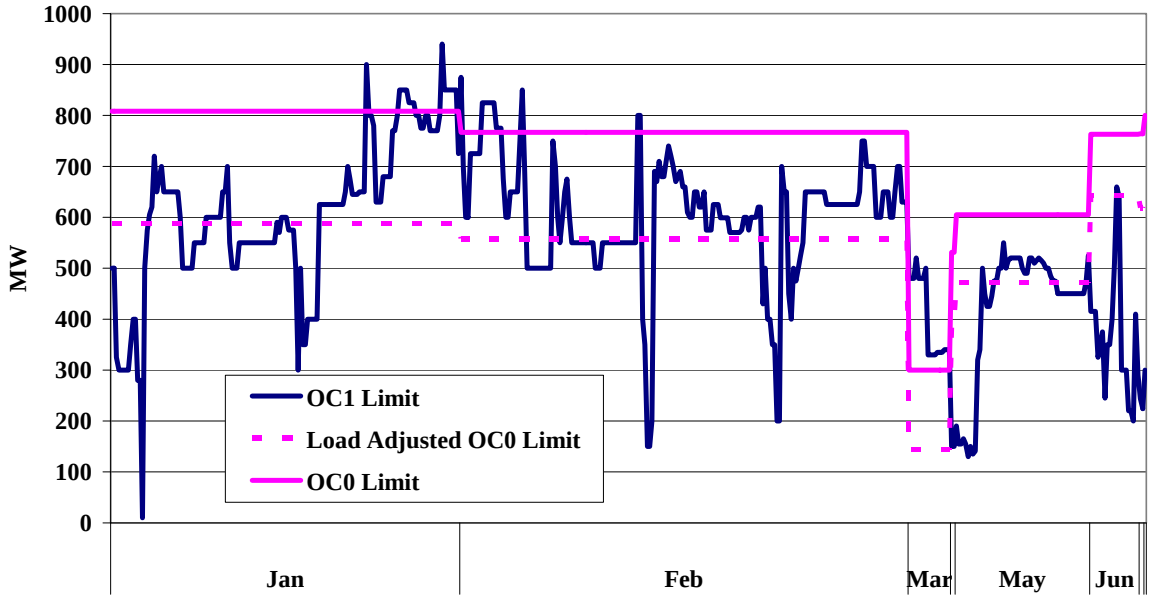
First, the topology of the network under actual conditions may differ from the topology of network in the monthly planning model. Second, because the GSFs of the individual generators in a zone can vary substantially from the zonal average shift factor used in the zonal model, changes in output from individual generators can change the flow over a

CSC, which requires the operators to adjust the transfer limit. Last, because the load within the zone is not distributed in the same manner as the generation, the effect of consumption in the zone (i.e., the load shift factor) will differ from the effect of the generation in the zone. Hence, changes in load levels, even if matched by changes in generation in the same zone, can change the flow over a CSC and cause the operator to change the OC 1 limit.

The load adjustment is the amount by which the flow over a CSC is altered due to the difference between load shift factor and generation shift factors. ERCOT calculates this load adjustment using monthly planning studies to establish the quantity of TCRs that it can sell. However, the actual load effect can change hourly, as described above.

The following four figures compare the OC 0 limit, the load-adjusted OC 0 limit, and OC 1 limit for each CSC in intervals when the CSC is constrained in real-time. Figure 4 shows this analysis for the South-to-North interface.

Figure 4: OC 0 Limits vs. OC 1 Limits During Congested Intervals
South-to-North Interface



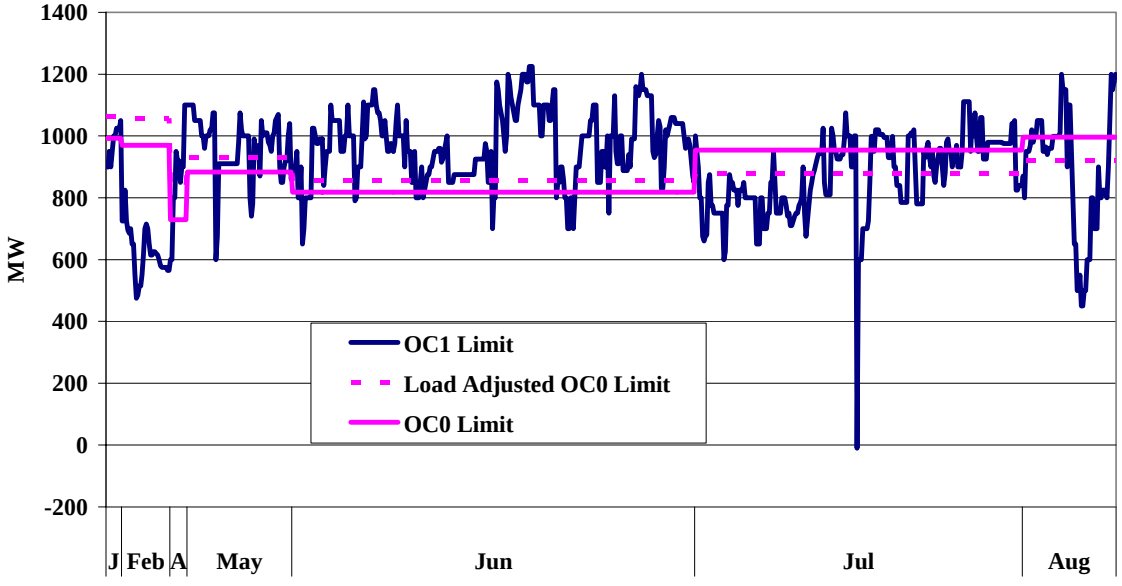
In Figure 4, the x-axis shows the congested intervals in chronological order from January through August 2004. The period from March to August was infrequently congested, while January and February experienced the most significant congestion. While the OC 0

(planning model) limits are relatively stable over the study period, the OC 1 (actual real-time) limits fluctuate substantially due to changes in system conditions. The ERCOT operators adjust the OC 1 limits in real-time in order to manage the power flows over the interface.

The load adjustment ranged from -119 MW to -221 MW over the study period, causing the load-adjusted OC 0 limits to be consistent with the average OC 1 limits. However, the analysis indicates that the OC 1 limits varied widely over the period. The load-adjusted OC 0 limits and OC 1 limits differed significantly in those intervals when the physical conditions on the system vary substantially from the conditions and simplifying modeling assumptions. The periods when the model diverges most substantially from reality raise serious concerns regarding the efficiency of interzonal congestion management.

Figure 5 compares the OC 0 limit, load-adjusted OC 0 limit, and OC 1 limit for the South-to-Houston CSC during constrained intervals.

Figure 5: OC 0 Limits vs. OC 1 Limits During Congested Intervals South-to-Houston Interface



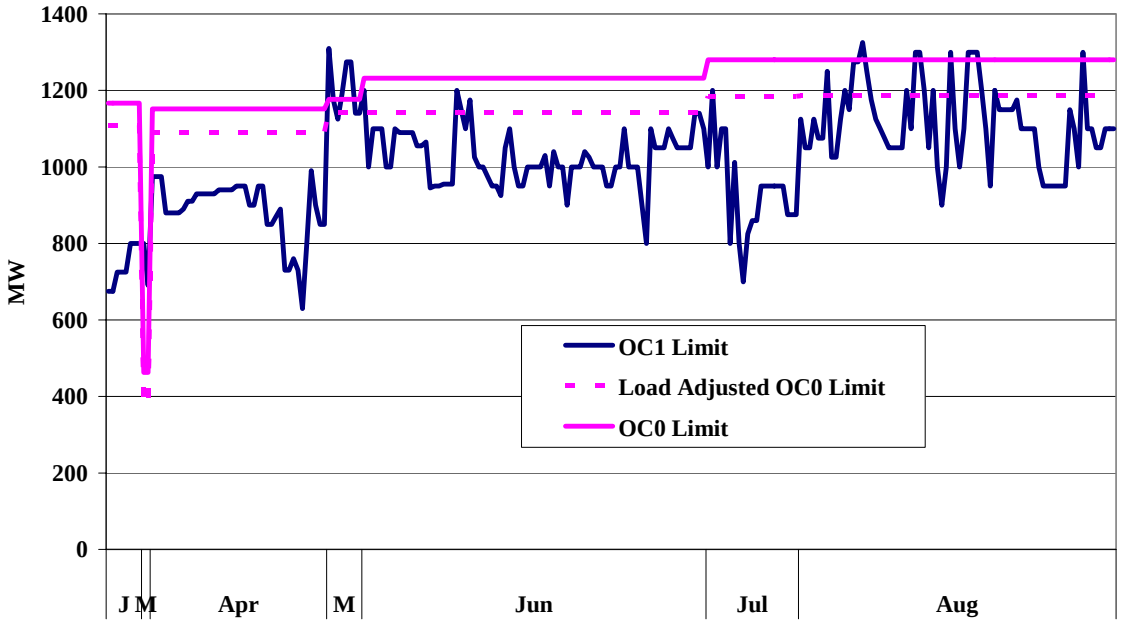
Approximately 40 percent of the intervals with congestion occurred during June when the load-adjusted OC 0 limit was 854 MW. During this period, the OC 1 limit fluctuated

between 650 MW and 1,225 MW, averaging 972 MW or 118 MW above the load-adjusted OC 0 limit.

During July, the average OC 1 limit was within 8 MW of the load-adjusted OC 0 limit. However, the OC 1 limit ranged from -10 MW to 1,111 MW during July. The occurrence of a -10 MW OC 1 limit indicates that the real-time model results differed significantly from physical conditions for a brief period.

Figure 6 shows the same analysis for the North-to-Houston CSC during constrained intervals. January, March, and May were only congested during a small number of intervals, while April and the summer months exhibited relatively frequent congestion.

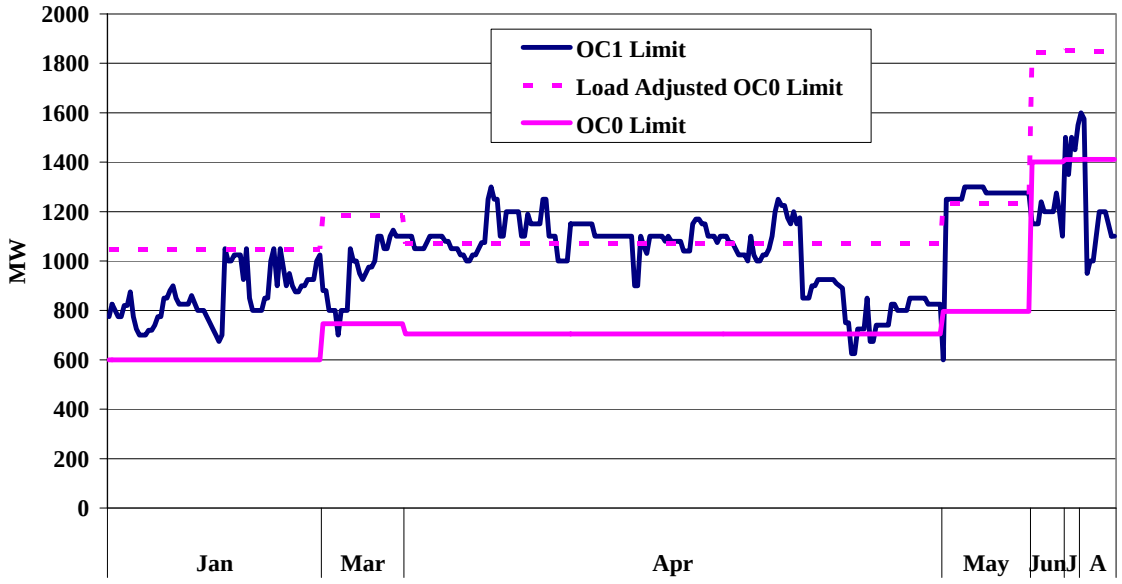
Figure 6: OC 0 Limits vs. OC 1 Limits During Congested Intervals
North-to-Houston Interface



Approximately 34 percent of the intervals with congestion occurred during June when the load-adjusted OC 0 limit was 1,143 MW. During this period, the OC 1 limit fluctuated between 800 MW and 1,200 MW, averaging 1,026 MW. While the OC 1 limit fluctuated throughout June, it was considerably less volatile than the OC 1 limits for the South-to-North and South-to-Houston interfaces. In general, the OC 1 limit was persistently lower than the load-adjusted OC 0 limit by 100 MW to 300 MW.

Figure 7 compares the OC 0 limit, the load-adjusted OC 0 limit, and OC 1 limit for the North-to-Northeast CSC during congested intervals. Most of the congested intervals occurred in January and August on this interface.

Figure 7: OC 0 Limits vs. OC 1 Limits During Congested Intervals North-to-Northeast Interface



Approximately one-half of the intervals with congestion occurred during April when the OC 0 limit was 1,070 MW. During this period, the OC 1 limit fluctuated between 625 MW and 1,300 MW, averaging 1,026 MW. The standard deviation of the OC 1 limit was 145 MW during April, largely due to a significant reduction in the OC 1 limit in the last portion of the month.

The analysis in the four figures above indicates that there have been moderate differences between the average of OC 1 limits used in real-time and the load-adjusted OC 0 limits calculated in the monthly planning studies for each of the CSCs. All of the OC 1 limits showed significant volatility, although the North-to-Houston and South-to-Houston limits were less volatile than the other limits for the other CSCs. The volatility of the OC 1 limits is not surprising given the design of the ERCOT market. The operators must make adjustments to ensure that the simplified zonal model effectively manages the flows and congestion over the CSCs as discussed above.

D. Assessment of Zonal Shift Factors

In this subsection we examine in more detail one of the most simplifying assumptions embedded in the current zonal market framework -- the assumption that all generators in a zone have the same effect on the flow over each CSC (i.e., the same shift factor).

When a resource generates electricity on the grid, the location of the resource and the load determines where the power actually flows. In particular, power will flow over the paths of least resistance from the point of injection (generator) to the point of withdrawal (load) based on the laws of physics.[6] Hence, resources in different locations will have different impacts on the flows of power over various transmission facilities. For the purposes of operating the balancing energy market and managing interzonal congestion, ERCOT assumes all resources within a zone have an effect on each CSC equal to the calculated zonal average shift factor.

This issue is important to analyze because differences between the actual resource-specific shift factors and the zonal average shift factors can result in inefficient congestion management and inaccurate zonal prices.[7] If the model recognized the relative effectiveness of the resources in affecting the flows over the CSC, it would redispatch different resources than are redispatched through the current market, which would generally result in lower zonal prices.

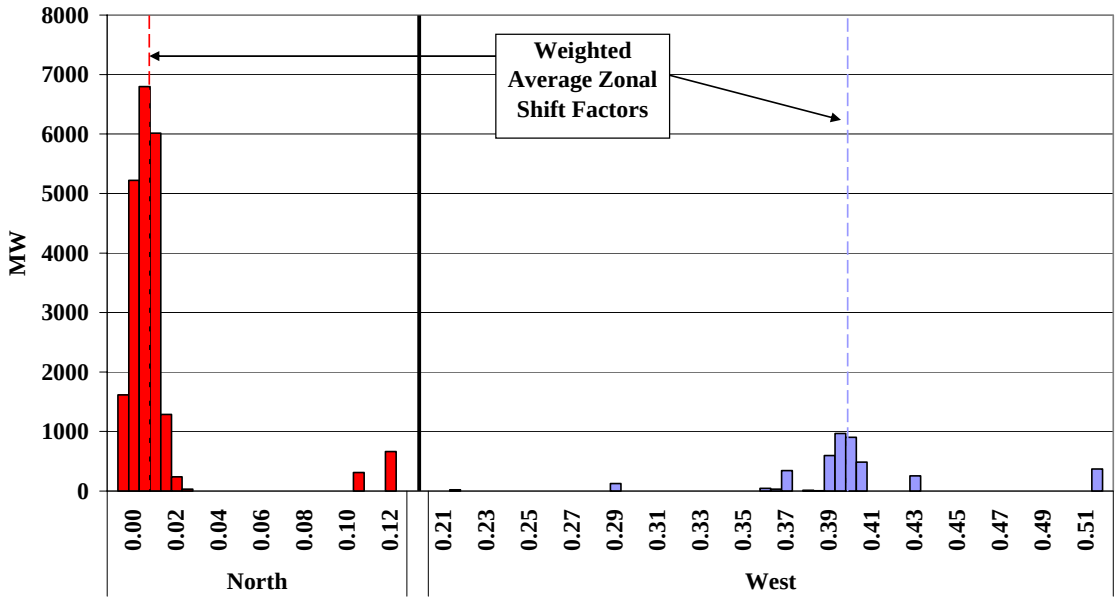
The zonal-average shift factors are based on annual planning studies and updated on a monthly basis. The planning cases calculate the shift factors of each of the generators within a particular zone under peak conditions. Generally, the monthly planning studies use cases with seasonal peak load conditions, dividing the total flow impact of the zone's output across the generation within the zone to calculate the zonal-average shift factor. Figure 8 through Figure 12 show the variation in impact of resources within a particular zone relative to the flow over each CSC in comparison to the zonal-average shift factor

6 In general, less power will flow over lower voltage, longer-distance paths.

7 Ross Baldick's previous work, "Shift factors in ERCOT congestion pricing," Ross Baldick, March 2003, http://www.ece.utexas.edu/~baldick/papers/shiftfactors.pdf, contains an analyses of how differences between zonal and local shift factors can impact efficiency of congestion management by the balancing market model.

calculated in the annual planning study. Figure 8 shows this analysis for the West-to-North CSC.

Figure 8: Distribution of Resource-Specific GSFs by Zone
West-to-North CSC -- 2004



While the resources in the other three zones impact flows over this interface, Figure 8 shows the shift factors only for resources in the North and West zones. The generation from the planning case is divided into 0.5 percent intervals and the value in the figure indicates the midpoint. For example, the column labeled 0.00 percent includes generation with resource-specific shift factors between -0.25 percent and 0.25 percent. The figure indicates that the resource-specific shift factors for resources in the North zone vary from -1 percent to 12 percent, while the zonal average shift factor is 1 percent. The resource-specific shift factors for resources in the West zone vary from 21 percent to 51 percent, compared to the zonal average of 40 percent.

The annual planning study uses a case simulating the system with a load level of approximately 69,000 MW. In this case, more than 22,000 MW would be on line and producing in the North zone. Most of the generation in the North zone has a slight positive impact on flows from the West to the North (i.e., generation that increases congestion from the West Zone to the North Zone when it increases production). Only 2,610 MW of generation has a negative shift factor, reducing flows over the West-to-

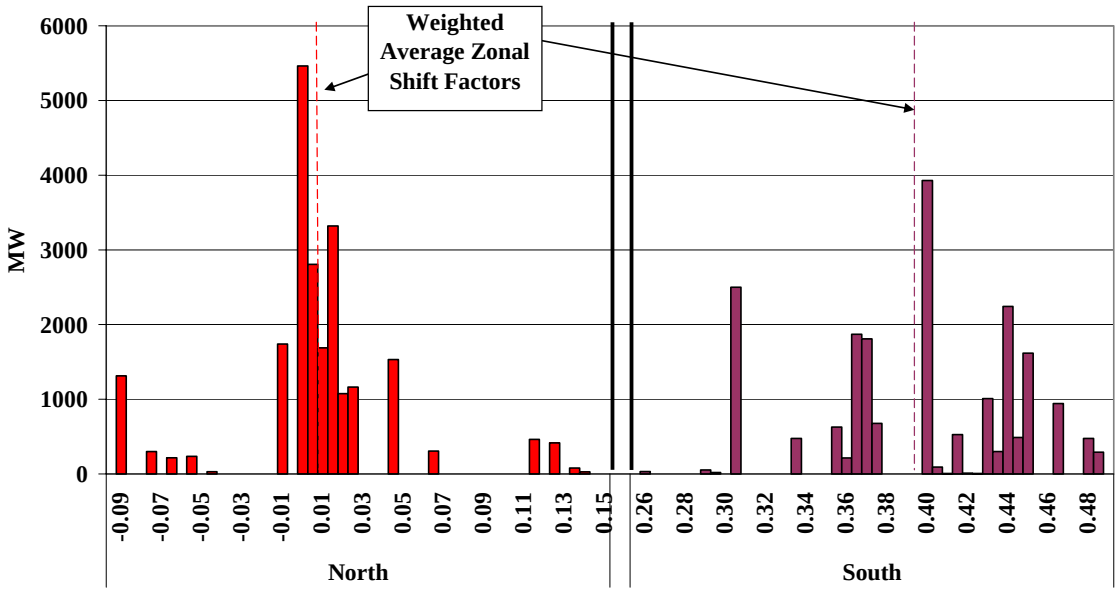
North CSC when it increases production. In addition, almost 1,000 MW of generation has a shift factor greater than 10 percent, substantially higher than the zonal average shift factor, which is significant because this 1,000 MW of generation has a larger impact on the flows over the CSC than all other 18,000 MW of generation in the North zone.

Likewise, the figure shows that the 4,159 MW of West zone generation online in the planning study has shift factors that vary, but not as significantly in percentage terms as those in the North zone. While the average shift factor is 40 percent, individual resources have shift factors as low as 21 percent and as high as 51 percent.

South-to-North CSC

Figure 9 shows the distribution of the shift factors for the South-to-North CSC. While the resources in the other three zones also impact flows over this interface, Figure 9 shows the detail of the shift factors for resources in the North and South zones because they have the largest effects on the South-to-North CSC. Like the prior figure, this figure shows the weighted average shift factor by zone as well as the distribution of the resource-specific shift factors.

Figure 9: Distribution of Resource-Specific GSFs by Zone
South-to-North CSC -- 2004



The South-to-North CSC includes a complex set of transmission elements that directly connect the South zone to the North zone. The figure indicates that the resource-specific shift factors for resources in the North zone vary from -9 percent to 15 percent, while the zonal average shift factor is 1 percent. Roughly half of the resources in the North zone actually cause power flows to *increase* over the South-to-North CSC when they increase their output.

Because GSFs vary so substantially within the North zone (and have different signs), they can have a significant effect on the modeled flows versus the real flows over the CSC. For example, reducing the output of generation in the North zone by 100 MW as part of a redispatch to manage congestion on the CSC could result in the flow decreasing 9 MW or increasing 14 MW depending on which units in the North zone respond to the redispatch.

This variation among generation units is all the more problematic because the choice of the responding unit is at the discretion of the QSE since ERCOT does not issue unit-specific instructions to solve interzonal congestion. The QSE does not have an incentive to dispatch the unit that has the most desirable impact on the constraint and, in fact, could have an economic incentive to dispatch the unit with the worst impact on the constraint in some cases. For example, a QSE can dispatch the resources with the lowest GSFs to maximize the flow on the CSC, which will require larger interzonal balancing energy deployments and result in higher prices in the import constrained zone.

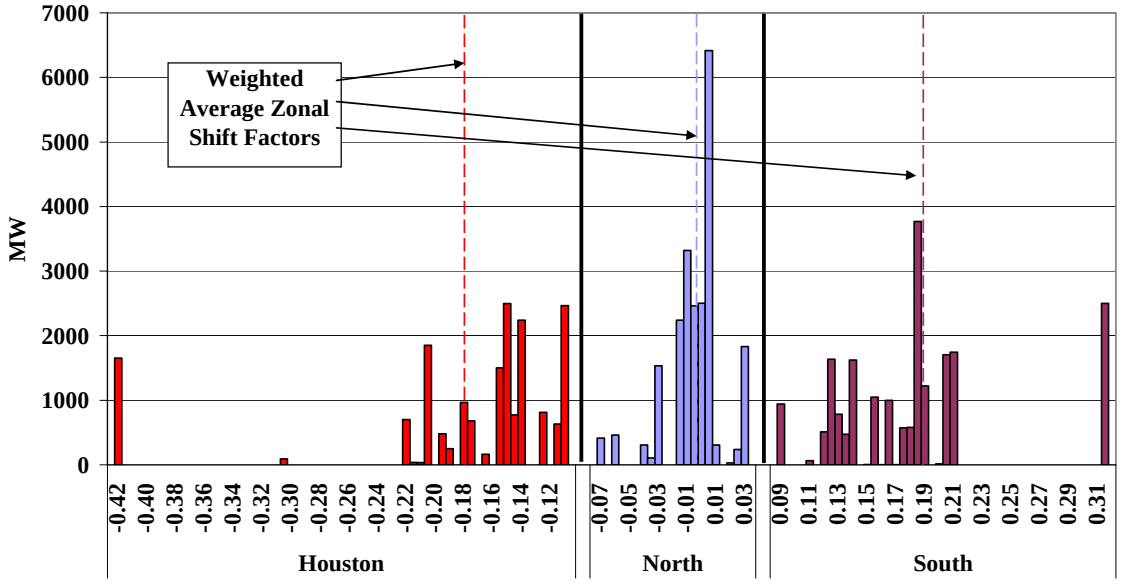
The shift factors for resources in the South zone vary from 26 percent to 49 percent and the zonal average shift factor is 39 percent. Figure 9 also shows the 20,000 MW of generation resources that were online in the South zone in the annual planning case. Of these, only 4,000 MW have shift factors within 1 percentage point of the zonal average shift factor. Almost 5,000 MW of resources have shift factors below 34 percent or above 45 percent.

South-to-Houston CSC

Figure 10 shows this shift factor analysis for the South-to-Houston CSC. In addition to showing the shift factors for units in the South and Houston zones, Figure 10 includes the

resources in the North zone because their resource-specific shift factors vary relatively widely for this CSC. Although the average zonal shift factor in the North zone is close to zero, making it unlikely that SPD will issue portfolio deployments in the North zone to manage the South-to-Houston CSC, some of the units in the North have a significant positive or negative effect on this CSC.

Figure 10: Distribution of Resource-Specific GSFs by Zone
South-to-Houston CSC -- 2004



The South-to-Houston CSC is comprised primarily of the 345 kV transmission lines that directly connect the South Texas Project in the South zone to the Dow plant in Houston. The figure indicates that the resource-specific shift factors for resources in Houston vary from -42 percent to -10 percent, the shift factors for resources in the North zone vary from -7 percent to 4 percent, and the shift factors for resources in the South zone vary between 9 percent and 32 percent. The figure also shows that the zonal-average shift factors for the Houston, North, and South zones are -18 percent, 0 percent, and 18 percent, respectively.

The shift factors for the resources in each of the zones vary substantially, making the effects of interzonal redispatch through the balancing energy market highly uncertain. Of the 18,000 MW of generation that was online in the Houston zone in the annual planning case, most of the resources have shift factors smaller than -15 percent and few resources

have shift factors close to the zonal average of -18 percent. Figure 10 also shows that resources in the South and North zones have resource-specific shift factors that vary significantly. For example, the resources in the North zone have shift factors that range from -7 percent to 4 percent.

North-to-Houston CSC

Figure 11 shows this analysis for the North-to-Houston CSC. In this case, the South zone generation is included in the figure because its resource-specific shift factors for the North-to-Houston CSC vary more widely than the shift factors in the source and sink zones (North and Houston). In addition, the zonal average shift factor for the South zone is large enough that it is likely the SPD will sometimes deploy balancing energy in the South zone to manage flows on the North-to-Houston CSC.

Figure 11: Distribution of Resource-Specific GSFs by Zone
North-to-Houston CSC -- 2004

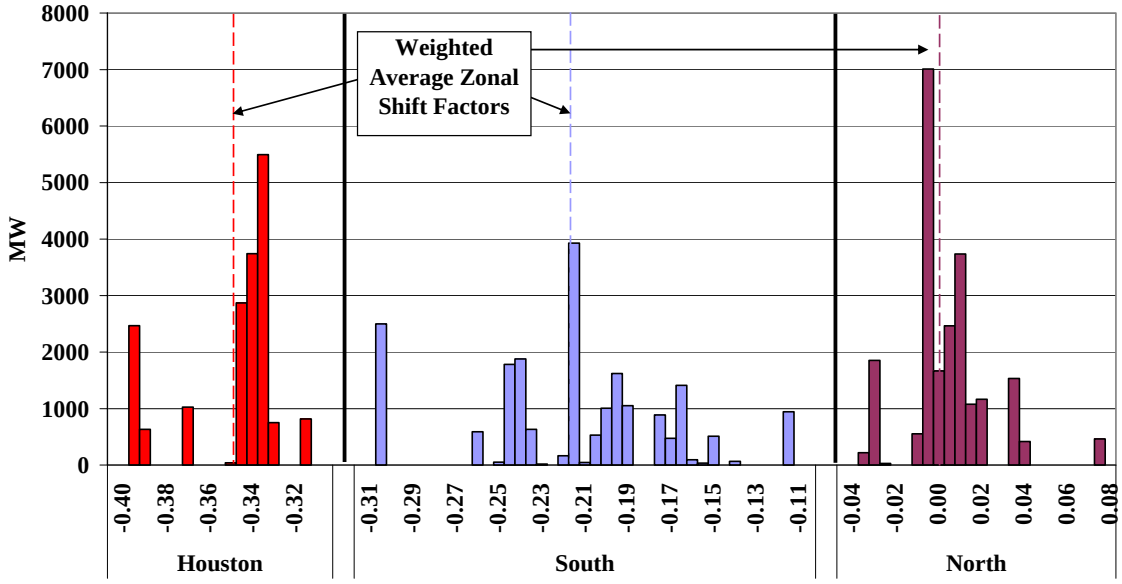


Figure 11 indicates that the resource-specific shift factors for resources in the Houston zone vary from -40 percent to -31 percent, a smaller dispersion than observed in the other zones. The shift factors for resources in the South zone vary from -31 percent to -11 percent, with a zonal average shift factor of -22 percent. The shift factors for resources in the North zone range from -4 percent to 8 percent and exhibit a zonal average of zero

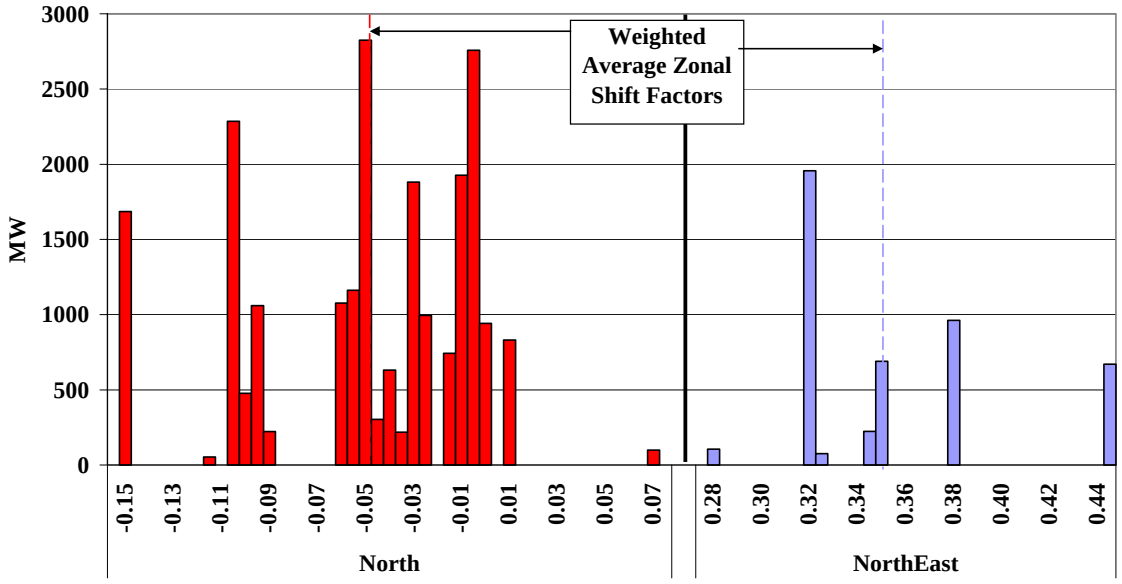
percent. As is the case for other CSCs, changing the output levels in the North zone can increase or decrease the flow over the North-to-Houston CSC.

The South zone has a very wide distribution of resource-specific shift factors. The South Texas Project has the largest impact with a shift factor of close to -30 percent while the Sandow resource has a shift factor of -11.6 percent. Note that a portfolio's shift factor can differ significantly from the zonal average. For example, the weighted average shift factor of TXU's portfolio in the South zone is -12.3 percent, versus the zonal average shift factor of -22 percent. Hence, the TXU generation in the South zone is much less effective at resolving congestion on this CSC than other generation in the South zone, information that the SPD model does not recognize when it redispatches zonal balancing energy to clear congestion on this CSC.

Northeast-to-North CSC

The final figure in this series shows the shift factors for resources in the newly created Northeast-to-North CSC.

Figure 12: Distribution of Resource-Specific GSFs by Zone
Northeast-to-North CSC -- 2004



The figure indicates that the resource-specific shift factors for resources in the North zone vary from -15 percent to 7 percent, while the shift factors for resources in the Northeast

zone vary from 28 percent to 45 percent. The zonal-average shift factor is -5 percent for the North zone and is 35 percent for the Northeast zone.

The balancing energy market model manages congestion on the Northeast-to-North interface primarily by increasing output in the North and decreasing output in the Northeast. However, when the portfolios in the North are incremented, the resulting impact on physical flows is uncertain because of the wide dispersion of shift factors. The problem caused by this inconsistency is particularly acute because the average shift factors of individual QSE portfolios also differ significantly.

For instance, FPL Energy Power Marketing is a QSE with over 1,600 MW of generation in the North zone that has a -15.1 percent shift factor, while Tractabel Energy Marketing is a QSE with over 900 MW that has an average shift factor of -0.9 percent. When the balancing energy market model evaluates portfolio offers from these QSEs, it does not take into account the substantially different impact that each will have on the flows over the Northeast-to-North interface.

Zonal Shift Factor Conclusions

Ultimately, the differences between the actual resource-specific shift factors and the zonal average shift factors result in inefficient congestion management and inaccurate zonal prices. If the model recognized the differences between units in affecting the flows over the CSC, ERCOT would redispatch different resources (or portfolios) than are redispatched through the current market. Improving these redispatch decisions would generally lower zonal prices.

The dispersion of generation shift factors also raises concerns that suppliers may have the opportunities to create congestion on CSCs without incurring its costs. To illustrate this point, we use the following hypothetical example:

- TXU's North zone portfolio includes the North Lake plant and the Mountain Creek plant. The North Lake plant has a 0.1 percent shift factor with respect to the Northeast-to-North interface, while the Mountain Lake plant has a -10.6 percent shift factor.

- When the balancing energy market model redispatches to resolve congestion on this interface, it uses a -5 percent zonal-average shift factor for all resources in the North zone.

- If the flow exceeds the interface limit, the balancing energy price and output levels would rise in the North zone and both would fall in the Northeast zone.

- If TXU's portfolio was struck for 100 MW, the model would estimate a resulting 5 MW reduction in flows over the CSC.

- However, TXU could actually create additional congestion by decreasing output from the Mountain Lake plant by 300 MW and increasing output from the North Lake plant by 400 MW. The net impact would be to increase the CSC flows by 32 MW.

- Before the subsequent interval, the ERCOT operators would likely decrease the OC 1 limit in the balancing energy market model to compensate for the inconsistencies between the flows calculated from zonal-average shift factors and the actual flows. This reduction would increase the difference in prices between the two zones.

The inefficient dispatch caused by inconsistent, if not contradictory, shift factors within a zone and the potential for gaming of these differences in shift factors cause us to conclude that ERCOT should move away from the current zonal market design in favor of a nodal market design as is currently under consideration.

The analysis in this subsection also indicates that the accuracy of the zonal shift factors could be improved. The extent to which the redispatch effects on the CSCs are overstated or understated can be reduced by calculating zonal average shift factors that reflect only the resources that are likely to be redispatched to manage interzonal congestion.

For instance, the South Texas Project ("STP"), a nuclear plant in the South zone, is notable because STP accounts for 2,500 MW of the supply used to calculate the zonal-average shift factor for the South zone. The shift factor for this resource is 30 percent on the South-to-North, almost 10 percent less than the zonal average shift factor for this CSC. If ERCOT were to remove the South Texas Project from the calculation of the zonal-average shift factor for the South-to-North CSC, the shift factor would more accurately predict the effects of a zonal redispatch since a nuclear plant will generally not

alter its output in response to portfolio balancing deployments from ERCOT. Likewise, removing this plant would reduce the zonal average shift factor for the South zone on the South-to-Houston CSC from 18 percent to 16 percent.

The same argument can be made regarding other baseload or cogeneration resources, such as the Comanche Peak plant in the North zone and the Dow plant in the Houston zone. If the Dow plant were removed from the calculation of the zonal-average shift factor for the South-to-Houston CSC, the average shift factor would change from -18 percent to -15.4 percent. Since a cogeneration plant will not usually be redispatched in response to a portfolio balancing energy deployment, the model will tend to over-estimate the impact of increasing output in Houston to manage congestion on the South-to-Houston CSC.

Hence, we recommend ERCOT consider modifying its methodology for calculating the zonal average shift factors to make them more accurate. This should improve the efficiency of the interzonal congestion management and zonal prices. However, this approach implies that the shift factors that ERCOT uses to calculate the base loadings on the CSCs may differ from the shift factors used to calculate the incremental effect of redispatch through zonal portfolio deployments. ERCOT can accommodate this inconsistency between these two sets of shift factors by automatically modifying the base loadings on each of the CSCs associated with output of the baseload generators that have been removed from the zonal shift factors and of the consumption by the load within the zone.

Effects of Load on the CSC Power Flows

The shift factors the balancing energy market model uses are based on a generation-weighted average of many individual resource-specific shift factors. The model also uses these shift factors to calculate the effect of load on flows over various interfaces.

It is also possible to calculate load-weighted shift factors. Unless load happens to be distributed identically to generation, the load-weighted shift factor will be higher or lower than the zonal shift factors used by the balancing energy model. While the differences

tend to be small on a per megawatt basis, the net effect can be large when applied to all load in a particular zone. Table 2 shows load-weighted and generation-weighted average shift factors from the annual planning study for all 25 zone-CSC combinations.

Table 2: Summary of Load and Generation Shift Factors

CSC	Zone	Load Weighted Shift Factor	Generation Weighted Shift Factor	Load v. Gen Difference	Summer Peak Load	Change in Flows at Summer Peak
West to North						
	West	37.2%	40.3%	-3.2%	4,104	-130
	North	0.7%	1.0%	-0.2%	24,950	-62
	South	3.4%	3.0%	0.4%	17,203	61
	Houston	1.8%	1.8%	0.0%	19,833	-3
	Northeast	0.7%	0.3%	0.3%	1,488	5
						-129
South to North						
	West	2.7%	2.3%	0.4%	4,104	18
	North	0.3%	0.6%	-0.3%	24,950	-81
	South	38.7%	39.1%	-0.4%	17,203	-69
	Houston	20.7%	20.9%	-0.2%	19,833	-38
	Northeast	0.5%	0.6%	-0.1%	1,488	-1
						-171
South to Houston						
	West	1.7%	1.5%	0.2%	4,104	9
	North	0.2%	-0.4%	0.6%	24,950	158
	South	16.4%	18.4%	-2.0%	17,203	-350
	Houston	-16.2%	-17.9%	1.7%	19,833	335
	Northeast	-0.2%	-0.3%	0.1%	1,488	1
						153
North to Houston						
	West	-2.0%	-1.8%	-0.3%	4,104	-10
	North	-0.4%	0.4%	-0.7%	24,950	-179
	South	-20.6%	-21.6%	1.0%	17,203	179
	Houston	-35.4%	-34.9%	-0.5%	19,833	-106
	Northeast	0.2%	0.3%	-0.1%	1,488	-1
						-117
Northeast to North						
	West	0.1%	0.2%	0.0%	4,104	-2
	North	-2.8%	-5.0%	2.2%	24,950	559
	South	-3.8%	-3.9%	0.1%	17,203	13
	Houston	-4.6%	-4.6%	0.0%	19,833	-1
	Northeast	25.3%	35.4%	-10.1%	1,488	-150
						418

In addition to the load-weighted and generation-weighted shift factors, Table 2 shows the shift factor difference, the summer peak load, and the impact this is likely to have on flows over each interface. For instance, the West zone has a load-weighted shift factor of 37.2 percent relative to the West-to-North CSC and a 40.3 percent generation-weighted shift factor. Therefore, when the model uses the generation-weighted shift factor, the calculation of flows from the West to the North will be biased downward due to the -3.2 percentage point difference. If the West zone load is modeled at 4,104 MW, as in the annual planning study, this will bias the flow calculation downwards by 130 MW. This "load adjustment" is calculated for each zone relative to a particular CSC, and then the total load adjustment is aggregated across the five zones for the CSC. The operators handle the bias by reducing the flow limit for the CSC in real-time by an amount roughly equal to the load adjustment. Since these effects are related to the load in each zone, they will generally fluctuate as the load fluctuates.

In most cases, the differences between load- and generation-weighted shift factors are relatively small. Twenty of the 25 zone-CSC combinations have differences of less than 1 percentage point. Moreover, many of the load adjustments are cancelled out by the load adjustments of other zones. For instance, the South-to-Houston interface flow calculation is significantly affected by the bias in the shift factor used for the Houston zone. Since the shift factor over-estimates the impact of the Houston load on the CSC, this will bias the flow calculation upwards by 335 MW. However, the South-to-Houston interface limit is also affected by the bias in the shift factor used for the South zone. The shift factor used by the model over-estimates the impact of South zone load on the CSC, which will bias the flow calculation downwards by 350 MW. In this case, the biases cancel each other and limit the effects of the load on the CSC limit.

In some cases, however, the load adjustments can have a significant effect on the zonal market results and cause some loads to incur higher congestion costs. To understand why, consider the following example.

- The North zone has a -2.8 percent load-weighted shift factor with respect to the Northeast-to-North interface, although the model treats load in the North zone as if its shift factor is -5.0 percent.

- Alternatively, the Northeast zone has a 25.3 percent load-weighted shift factor with respect to the Northeast-to-North interface, but the model uses a 35.4 percent shift factor for the zone, which results in a load adjustment of more than 400 MW in the planning study.

- In other words, the real dispatch of generation to meet load in ERCOT during peak conditions will load the Northeast-to-North interface by 418 MW less than the balancing energy market model will detect, resulting in an increase in the CSC transfer limit in the model.

If the operator could perfectly adjust the OC 1 limit (in this example increasing it by 418 MW), then the zonal prices would not be affected. However, the model would indicate that interzonal schedules and deployments are resulting in more flow over the CSC than they are in reality. Therefore, when the CSC constraint is binding, more congestion costs will be collected from the participants than would be collected if the system were modeled accurately. This issue exists even if the OC 1 limits are adjusted perfectly. In reality, it is a very challenging task for the operator to set OC 1 limits that accurately reflect the difference between physical flow and actual flow. To understand this process, consider the following example:

- Assume the physical limit of a CSC is 1,000 MW. If the physical flow rises to this level and SPD calculates flows of 600 MW, the operator must adjust the interface limit in SPD to 600 MW in order to prevent additional physical flow.

- If the physical flow rises by 50 MW in a subsequent interval, the operator will need to make a similar downward adjustment to the SPD limit. However, since the flow calculated by SPD is incorrect, the operator will not know whether a 50 MW adjustment to the SPD limit will cause too little or too much redispatch.

As this example shows, setting the SPD limit is a trial-and-error process for the operator because of the inaccuracies in the model. The adjustments to the CSC limits that must be made by the ERCOT operators are dependent on the accuracy of the assumptions in the model and the resulting flows over the CSCs produced by the model. The next subsection provides an evaluation of the accuracy of these flows.

E. Accuracy of Modeled Flows

The zonal market design that has been adopted in Texas makes a number of simplifying assumptions discussed above. Simplifying assumptions do not raise concerns unless they compromise the accuracy of the model. Significant inaccuracies can result in inefficient

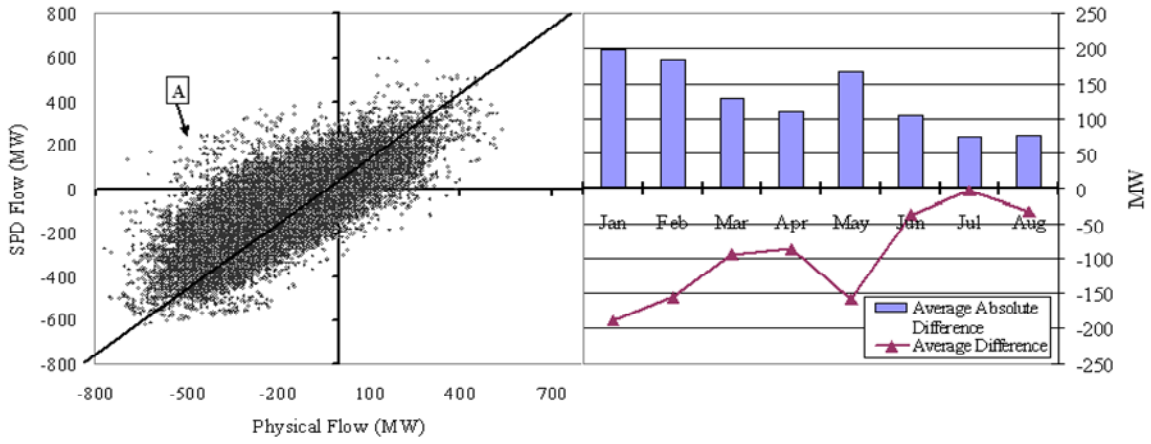
congestion management and higher costs to loads. Hence, we evaluate the accuracy of the ERCOT model in this sub-section by measuring the differences between the modeled flows and the actual flows over each CSC.

West-to-North Interface

We begin by showing the differences between actual physical flows and flows that SPD calculated across the West-to-North interface in Figure 13. The panel on the left is a scatter plot showing the SPD flow on the y-axis versus the actual flow on the x-axis during every interval from January through August 2004. The scatter plot also includes a 45 degree line showing where points would lie if the SPD flows equaled the actual flows.

The panel on the right chart shows two measures of the differences between SPD flows and actual flows on a monthly basis. The bar is the average difference (regardless of direction) between the SPD flow and actual flow across all intervals, which we calculate by averaging the absolute value of the difference in each interval. The line shows the average physical flow across all intervals minus the average SPD flow across all intervals. The line captures systematic differences where one quantity is consistently larger than the other. If all differences between modeled flows and actual flows were random, we would expect the line to be close to zero while the bars could be large.

Figure 13: Actual Flows vs. SPD Flows on the West-to-North Interface January to August 2004



In the left panel of Figure 13, the points in the figure trend upward with a slope that is slightly less than the 45 degree line. While there is a positive correlation between SPD flows and physical flows, the figure shows that there are a large number of intervals where the SPD flows and physical flows differ substantially.

The figure shows that the modeled flow and physical flow sometimes run in opposite directions. For example, point A in the figure is an instance when the SPD flow is approximately 200 MW from West-to-North while the actual physical flow is more than 500 MW *from North to West.* Overall, the physical flows and SPD flows run in opposite directions in almost 16 percent of the intervals.

In addition, the left panel shows that a substantial majority of intervals show flows moving from the North zone to the West zone. ERCOT plans to define a North to West CSC in 2005 in order to manage these flows out of the North zone.

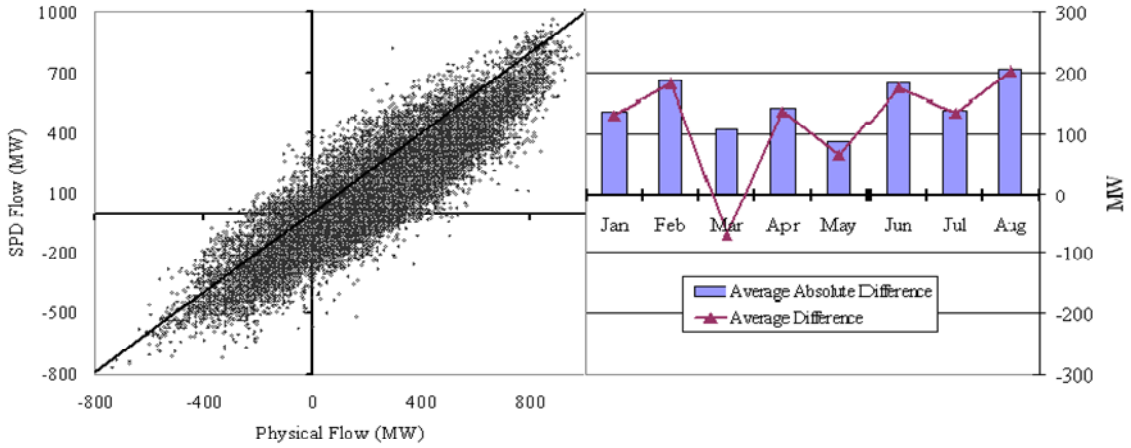
Based on the load adjustment highlighted in Table 2, we would expect SPD flows to be less than physical flows by approximately 129 MW during the summer peak load conditions. While this load adjustment is proportional to load, it can also change when the distribution of load between zones changes. Nonetheless, the load adjustment indicates that the physical flow should be *greater* than the SPD flows. However, the right panel in Figure 13 shows that the physical flows have been *less than* the SPD flows on average (the line in the figure), ranging from -30 MW to -200 MW on a monthly average basis. The bars show that the average absolute difference in the flows ranged from 75 MW in July to 200 MW in January.

These differences are significant given that the physical limit on the West-to-North interface is close to 500 MW on average. Hence, the monthly average difference ranges from less than 10 percent to 40 percent of the true interface limit. This range is much wider based on the interval level results. However, the inaccuracy of flow calculations has not led to significant inefficiencies during the study period because there has been virtually no congestion on the West-to-North interface.

South-to-North Interface

Figure 14 is similar to Figure 13. It shows two panels summarizing differences between actual physical flows and SPD calculated flows across the South-to-North interface.

Figure 14: Actual Flows vs. SPD Flows on the South-to-North Interface January to August 2004



Based on the load adjustment highlighted in Table 2, we would expect SPD flows to be less than physical flows by approximately 171 MW during the summer peak load conditions. This is generally consistent with the monthly average differences shown in the right panel of the figure and the fact that most of the points lie below the 45 degree line in the left panel of the figure.

Although the consistency of the flows is greater for this CSC, the scatter plot shows many intervals where the SPD flows and physical flows differ by a significant margin. In fact, the SPD flows and physical flows run in opposite directions in almost 20 percent of the intervals.

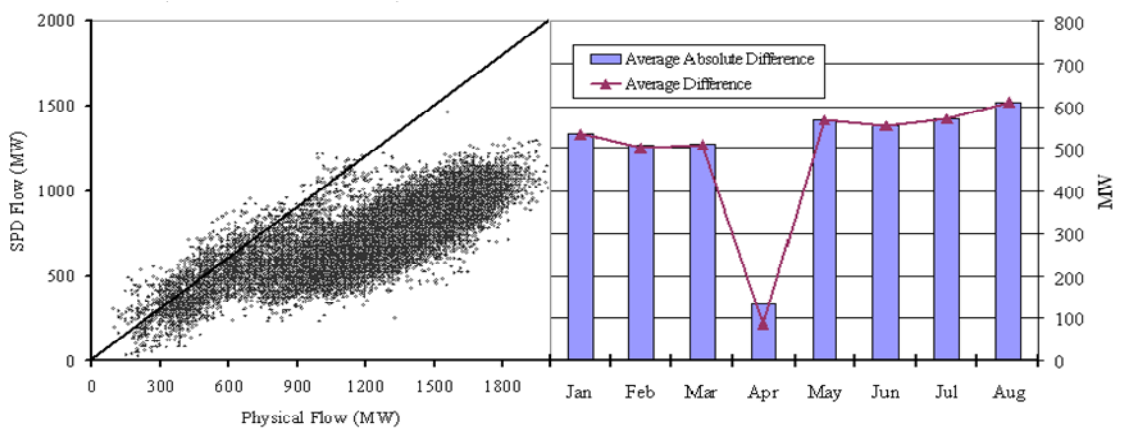
The physical limit of the interface averaged 753 MW during the study period. Hence, the monthly average differences between SPD flows and physical flows ranged from close to 10 percent to almost 30 percent of the physical limit of the CSC. When the interface becomes congested and the SPD flows differ from the physical flows, the operator must

adjust the limit of the interface for SPD to recognize the congestion as discussed in the previous subsection.

South-to-Houston Interface

Figure 15 includes two panels summarizing differences between actual physical flows and SPD calculated flows across the South-to-Houston interface.

Figure 15: Actual Flows vs. SPD Flows on the South-to-Houston Interface January to August 2004



Based on the load adjustment highlighted in Table 2, we expect physical flows to be *less than* SPD flows by approximately 153 MW. The bars show that in every month other than April, the average physical flows were actually *higher* than the SPD load by an average difference of 500 to 600 MW. Therefore, the total difference between the load-adjusted SPD flows and the physical flows was in the 700 MW range for most of the study period. This difference is considerable relative to the total physical capacity of the interface, which averaged 1,853 MW during the study period. The results for April were anomalous compared with the other seven months because the average SPD flow was higher than the average physical flow by only 130 MW.

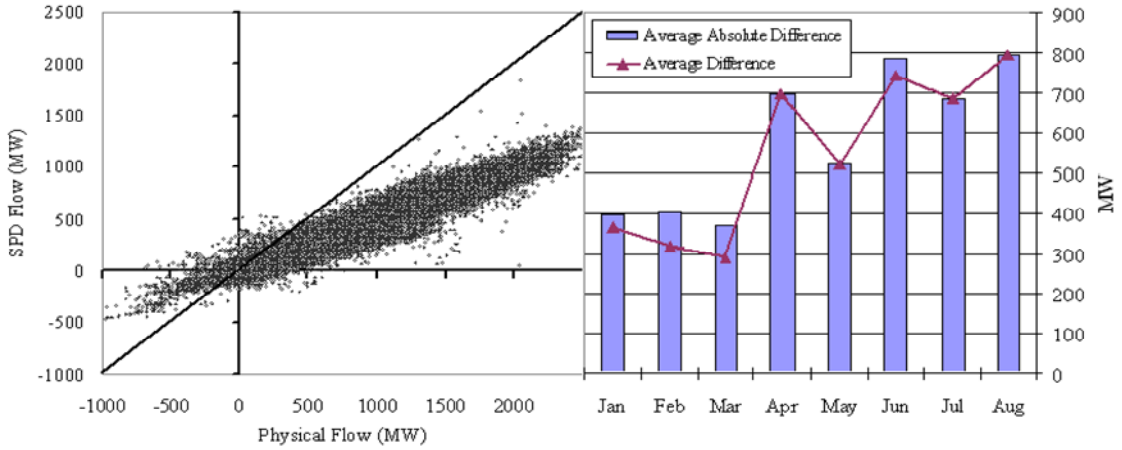
The scatter plot in Figure 15 shows that when flow levels are relatively low, SPD flows and physical flows tend to be closer than at higher flow levels. As physical flows grow, the inconsistency between the physical flows and SPD flows increases. The trends in this data indicate that for every 3 MW that physical flows increase, there is an average

increase of 2 MW in SPD flows. This requires that the operators continually adjust the OC 1 limit in the model as the physical flows change to maintain an appropriate SPD limit for the South-to-Houston interface.

North-to-Houston Interface

Figure 16 summarizes the differences between actual physical flows and SPD-calculated flows across the North-to-Houston interface. Like the previous figures in this analysis, the panel on the left is a scatter plot showing the SPD flow versus the actual flow, while the panel on the right is a bar/line chart measuring the differences between SPD flows and actual flows on a monthly basis.

Figure 16: Actual Flows vs. SPD Flows on the North-to-Houston Interface January to August 2004



The bars in Figure 16 show that the average difference in each month of the study period ranged from 380 MW to 800 MW. The average physical flow was higher by 300 MW to 800 MW during the study period. The difference between SPD flows and physical flows far exceeds the anticipated load adjustment calculation, since the estimated load adjustment predicts that physical flows would be higher by 117 MW. These differences between the physical flows and SPD flows are substantial given that the physical limit for this CSC averages close to 1,200 MW. Hence, the monthly average differences in the flows range from one quarter to two-thirds of the physical capability of the CSC.

The scatter plot shows flows in a tight distribution increasing at roughly one-half the slope of the 45 degree line. As with the South-to-Houston interface, the difference between the SPD flows and physical flows over this interface grow large at high flow levels when the interface is most likely to be congested. Like the prior CSC, this implies that operators must continually modify the OC 1 limits for this CSC to ensure the model remains well-calibrated to reality.

Northeast-to-North Interface

Our final analysis in this series of figures is shown in Figure 17, which summarizes differences between actual physical flows and SPD-calculated flows across the Northeast-to-North interface. The panel on the left is a scatter plot showing the SPD flow versus the actual flow, while the panel on the right is a bar/line chart measuring the differences between SPD flows and actual flows on a monthly basis.

Figure 17: Actual Flows vs. SPD Flows on the Northeast-to-North Interface January to August 2004

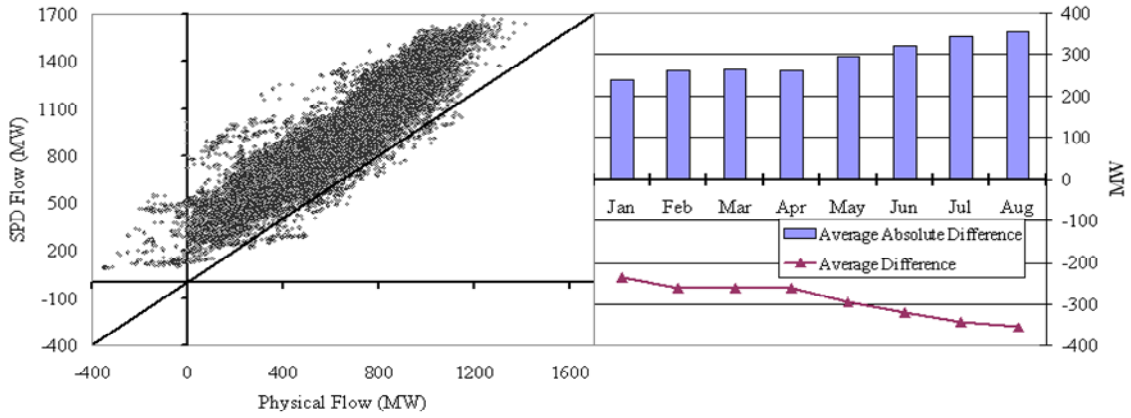


Figure 17 show that the SPD flow exceeded the physical flow in each month of the study period by average amounts ranging from 230 MW to 350 MW. The estimated load adjustment shown in Table 2 was 418 MW. Hence, the differences observed on this CSC are generally consistent with the load adjustment, although the points are dispersed widely.

Conclusions Regarding Accuracy of the SPD Flows

Figure 13 through Figure 17 indicate that there are persistent differences between the SPD flows and actual flows on the ERCOT CSCs. The load adjustments estimated for each CSC help to explain some of these differences. However, the differences between the SPD flows and the actual flows on each of the CSCs are very large in many intervals, far exceeding the load adjustments. On three of the CSCs, there are a significant number of intervals when the SPD flows and physical flows run in opposite directions.

These results raise a number of concerns regarding the accuracy of the zonal modeling framework. First, simplifying assumptions that result in inaccurate modeling of the network flows place a significant burden on the ERCOT operators to adjust the OC 1 limits continually to ensure the balancing energy market does not allow physical flows to exceed the physical limits of the CSCs. This explains the volatility of the OC 1 limits shown in the prior subsection.

Second, these results raise significant concerns about the efficiency of balancing energy deployments that ERCOT uses to manage interzonal congestion. Inefficient balancing energy deployments resulting from inaccurate zonal shift factors can distort the zonal balancing energy prices when the CSCs are binding and cause inefficient balancing energy deployments in different zones. Lastly, the differences in zonal energy prices and balancing energy deployments will affect the amount of congestion revenue collected from market participants. Therefore, it is important to make the simplifying assumptions as accurate as possible and to avoid over-simplifying the market framework by failing to define the optimal number of zones and CSCs.

F. Redispatch Analysis

In the subsections above, we have described and analyzed the zonal balancing energy market used to manage interzonal congestion under the ERCOT market rules and protocols. Much of this analysis has focused on the differences between the simplifications in the zonal framework relative to the actual operation of the system. The most important simplification is the use of portfolio supply offers and zonal shift factors to model the flows over the CSCs and manage interzonal congestion, which is not

efficient because the market is not able to select the most effective and economic resources within a zone to manage the congestion. This subsection analyzes the significance of this simplification by evaluating how much it affects the quantity of generation that must be redispatched to manage the interzonal congestion.

When the resource-specific shift factors vary substantially, the optimal unit-specific redispatch in a nodal market to manage the congestion can greatly differ from a portfolio-based redispatch in a zonal market. To understand why, consider the following example:

- Suppose that the flows over a particular CSC are over the limit by one MW and two identical resources are available to be dispatched up on the import-side:
  - Unit A with a shift factor of -0.4 on the CSC. The flow could be brought down to the interface limit by increasing output from Unit A by 2.5 MW (1 MW reduction = -0.4 * 2.5 MW).
  - Unit B with a shift factor of -0.1. The constraint could be resolved with a 10 MW increase from Unit B (1 MW reduction = -0.1 * 10 MW).
- Although Unit A is clearly more effective, the balancing energy market will redispatch both units (on a portfolio basis) as if their shift factors were equal to the zonal average. If the zonal average shift factor were -0.25, the balancing energy market would redispatch 4 MW.
  - If the more effective unit happened to respond, the flow would be reduced by 2 MW, twice the necessary amount.
  - If the less effective unit happened to respond, the flow would be reduced by only 0.4 MW, achieving less than one-half of the required relief.

One advantage to using resource-specific offers rather than portfolio offers to resolve energy imbalances in real-time energy markets is that resource-specific redispatch uses the most cost-effective units to manage all transmission constraints on the system. This subsection provides a comparative analysis of redispatch using the portfolio-based offers that actually occurred during the study period versus the economic redispatch that would have been possible with resource-specific offers. We would note that the Texas nodal markets currently under consideration would achieve these efficiencies because it would utilize resource-specific offers and shift factors.

To conduct this analysis, we measured the quantities of energy that ERCOT redispatched to reduce the flow on the CSCs in all constrained intervals. We used a simulation model to accomplish the same reduction in flow by redispatching units in the same intervals based on resource-specific shift factors. We redispatched only units with available incremental and decremental capacity in our simulation. Using this methodology, we conducted the following two redispatch scenarios:

- **Minimum redispatch:** In this scenario, the most effective generating units at relieving flow on the CSC are selected (based on resource-specific shift factors), regardless of their cost.

- **Economic redispatch:** In this scenario, the resource-specific offers together with unit-specific shift factors were used to choose the most economic resources to redispatch to relieve the flow on the CSC.[8]

While the minimum redispatch scenario identifies the redispatch alternative that would rely on the least amount of redispatch, the economic redispatch scenario is more representative of what would occur in a market with resource-specific offers. Economic dispatch generally results in a higher quantity of redispatch because less effective generators (i.e., smaller impact on the CSC) may be redispatched in lieu of higher-cost resources that are more effective at relieving the CSC.

We calculate a statistic for this evaluation that we refer to as a "redispatch ratio". We calculate this ratio by dividing the simulated redispatch quantity by the actual redispatch quantity. Lower ratios indicate that smaller quantities of redispatch would have been required to achieve the necessary relief on the given CSC. For example, a redispatch ratio of 66 percent indicates that the desired CSC relief could have been provided by redispatching only two-thirds of the quantity of generation that was actually redispatched. The results of this analysis are presented in Table 3 by CSC.

[8] The simulations were run using a linear programming model that minimizes total production costs using generator shift factors, the flows and limits and all active constraints, and resource-specific offers are based on the formula-based premiums for OOME dispatch in the ERCOT protocols. Generator commitments were held constant.

Table 3: Analysis of Redispatch Quantities

			Minimum Redispatch		Economic Redispatch	
CSC	Number of Intervals	Actual Redispatch	Redispatch Amount	Redispatch Ratio	Redispatch Amount	Redispatch Ratio
NorthEast-to-North	316	358	197	55%	221	62%
South-to-North	353	301	181	60%	199	66%
North-to-Houston	186	462	301	65%	358	77%
South-to-Houston	693	463	200	43%	356	77%
All CSCs				51%		72%

The results show that the average redispatch ratio for all of the CSCs in the minimum redispatch scenario was 51 percent. This indicates that based solely on the shift factors of the individual resources, the current zonal market framework resulted in close to double the quantity of redispatch during the study period as would have occurred by redispatching individual units to manage the same interzonal congestion. At the individual CSC level, the redispatch ratios in the minimum redispatch scenario ranged from 43 percent to 65 percent.

The average redispatch ratios in the economic redispatch scenario are higher as expected. The average ratio in this case was 72 percent for all of the CSCs. For the individual CSCs, redispatch ratios ranged from 62 percent to 77 percent. The 62 percent redispatch ratio for the Northeast-to-North interface indicates that for every 62 MW that would be redispatched optimally to manage congestion on this CSC, the zonal market will redispatch 100 MW – an increase in the quantity of generation redispatched of more than 60 percent. In fact, these results show that, all things equal, the quantity of generation redispatched under the current zonal market is 30 to 60 percent higher than the quantities that would be redispatched economically under a market utilizing resource-specific offer curves and shift factors.

These results likely understate the effects of moving to a full nodal electricity market as is currently under consideration in Texas. Not only would resource-specific characteristics be considered in the dispatch of generation, but also in the commitment of the generation. Hence, units that may not be committed under the current market design

due to their costs could be committed economically under a nodal market design because they are particularly effective at managing the flow on a constrained transmission facility.

G. Interzonal Congestion Conclusions and Recommendations

The analysis in this section leads to a number of conclusions regarding the current ERCOT market and interzonal congestion management procedures:

- After direct assignment of congestion rents on CSCs began in February 2002, interzonal congestion has been relatively infrequent and the costs have been modest;

- The zonal market framework used to manage interzonal congestion makes important simplifying assumptions that may not reasonably reflect operating conditions; and

- These simplifying assumptions can significantly affect the outcomes of the markets and can result in inefficient dispatch of the generating resources.

These shortcomings would have to be addressed to eliminate the existing market and operational inefficiencies. Unfortunately, these issues are inherent in the zonal market design so the potential for significant improvement is relatively limited. Other than implementing a nodal market design that would use resource-specific offers to efficiently manage all network constraints, the primary means to improve market performance is to optimize the designation of the zones and CSCs. This involves periodically evaluating the definition of the zones to ensure that the zones are designated to minimize the effects of the simplifying zonal assumptions. However, this must be balanced against preserving the stability of the market. Increasing the number of zones can affect the existing bilateral contracts and the liquidity of the forward market.

The process for creating new zones and CSCs is set forth in the ERCOT Protocols (section 7.2.1). The Congestion Management Working Group (CMWG) determines candidates for new zone or CSC definitions based on transmission pathways that are managed using significant amounts of Reliability Must Run resources, out-of-merit

capacity, and out-of-merit energy. For transmission pathways that result in significant uplift, the CMWG can set an uplift cost threshold that, if exceeded, makes the pathway a candidate for CSC treatment. ERCOT studies candidates for CSC treatment, as well as any candidates for new zones. The ERCOT Board, relying upon recommendations by ERCOT subcommittees, makes the final determination about whether to approve a change if the candidate zones and/or CSCs would reduce uplift costs and allow for a sufficiently competitive market (i.e. that a Market Solution exists).[9] We believe that this process could be improved in at least three ways:

- ERCOT should have the ability to identify proposed changes in the designation of CSCs and zones since they will likely be better informed regarding the patterns of congestion that have been occurring.

- The criteria used by ERCOT to implement new zones or CSCs should be amended to consider efficiency improvements beyond simply uplift quantities. For instance, creating new CSCs or zones could significantly improve the efficiency of the balancing energy deployments used to manage interzonal congestion on existing CSCs.

- Market power is important, but the minimum criteria for competitiveness set forth in the Protocols (i.e. that a Market Solution exists) is too strong. With appropriate provisions for dealing with local market power, the benefits of new CSCs and/or zones may outweigh concerns about market power.

The first candidate zone that we recommend ERCOT evaluate based on the analysis in this report and our prior 2003 State of the Market report is the Dallas/Ft. Worth zone. This would allow a large share of the congestion that is currently managed with out-of-merit energy, out-of-merit capacity, and Reliability Must Run resources to be priced more efficiently and transparently. We understand that there would be significant issues to consider in forming such a zone, including the effect on current bilateral contracts, the need for measures to effectively mitigate market power in the area, and the equity implications of such a change. ERCOT would need to evaluate these factors.

[9] A Market Solution is a criteria used to determine that the market is competitive. A Market Solution exists whenever three or more unaffiliated suppliers were capable of relieving a local constraint.

II. LOCAL CONGESTION MANAGEMENT

The ERCOT zonal market manages interzonal congestion on CSCs and Closely Related Elements ("CREs") using zonal shift factors. ERCOT considers all other congestion local and manages it by redispatching specific resources. ERCOT takes the following actions to manage local congestion:

- Manually committing a resource out-of-market that will help relieve the local congestion, known as out-of-merit commitment or ("OOMC");
- Committing or dispatching a reliability must-run resource ("RMR");
- Manually redispatching a specific unit or portfolio for out-of-merit energy ("manual OOME"); or
- Redispatching specific units through SPD by activating local constraints in the model ("OOME").[10]

In all cases, ERCOT performs resource-specific dispatch separately from portfolio dispatch by the balancing energy market model. ERCOT now manages most local congestion through SPD, which is the primary focus of this section.

A. Summary of Congestion Costs during the Study Period

In this subsection, we summarize local congestion and local reliability expenses that ERCOT incurs when resolving transmission constraints and reliability requirements that the current zonal market does not recognize, including constraints that ERCOT has not defined as part of an existing CSC. ERCOT manages local congestion by redispatching specific resources. When QSEs collectively do not commit the sufficient capacity to meet localized energy and reliability requirements, ERCOT commits additional resources to meet these requirements.

[10] These deployments are frequently referred to as local balancing energy, however they are currently settled in the same manner as the manual OOME. For the purposes of this report, therefore, we refer to all local deployments as OOME.

When ERCOT dispatches a unit out of merit for energy (either OOME Up or OOME Down), it pays the unit for a quantity equal to the difference between the scheduled output based on the unit's resource plan and the OOME instruction from ERCOT.[11] The payment/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, a resource with an OOME Up payment amount of $60/MWh that receives an OOME Up instruction when the balancing energy price is $35/MWh will receive an OOME Up payment of $25/MWh ($60-$35).[12]

For OOME Down, the Protocols establish an avoided cost amount 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/MWh receives an OOME Down instruction when the balancing energy price is $35/MWh, then ERCOT will make an OOME Down payment of $20/MWh. The OOME Down payment is a "make-whole" payment for the lost profit to the unit for the energy scheduled but not deployed.

ERCOT pays a unit providing capacity under an OOMC instruction according to a pre-determined formula defined in the ERCOT Protocols based on the type and size of the unit, natural gas prices, the duration of commitment, the balancing energy market clearing price, and whether the unit incurred start-up costs. Owners of a resource receiving an OOMC instruction from ERCOT must offer the energy between the resource's minimum and maximum capability into the balancing energy market.

Finally, units that ERCOT contracts to provide RMR service to ERCOT receive compensation for start-up costs, energy costs, and also receive a standby fee. The analyses in this section separate RMR uplift into two categories: (a) capacity costs, which

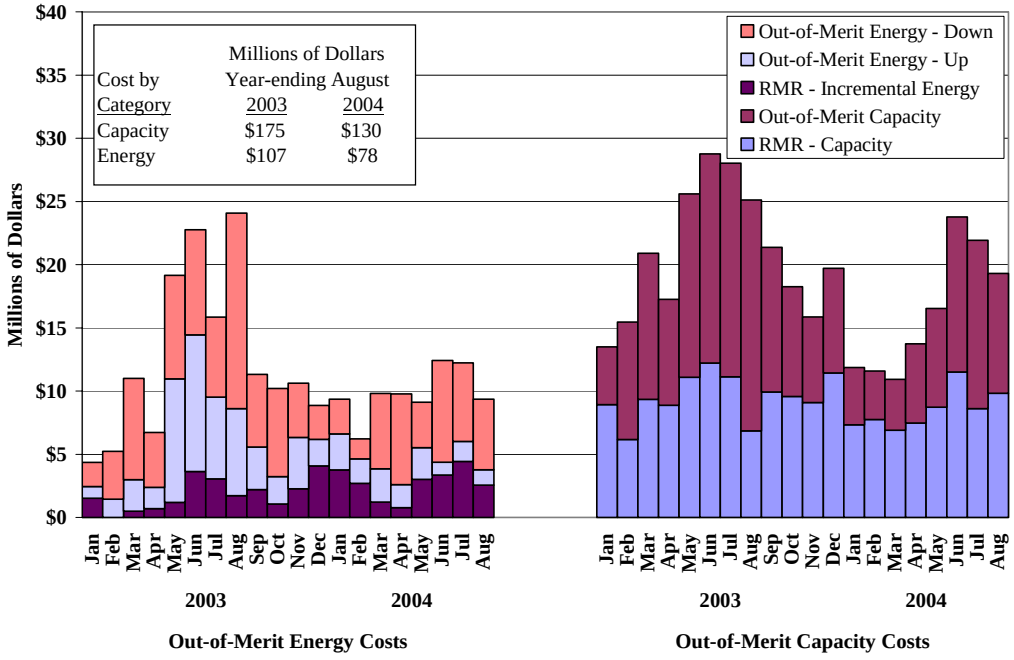
[11] This assumes all the parties will follow the OOME instructions without any variance. In reality, when there is a difference between the OOME instructions and the unit's actual response, ERCOT considers the metered output for the instructed resources for settlement purposes.

[12] OOME payments cannot be negative.

include start-up costs, standby fees, and energy costs up to the minimum dispatch level, and (b) incremental energy costs, which are the costs associated with output above the minimum dispatch level.

Figure 18 shows each of the five categories of uplift costs by month for 2003 as well as the first eight months of 2004. The left side shows costs of OOME (Up and Down) and incremental energy from RMR units, while the right side shows the net costs of RMR units and OOMC units. The net cost for RMR units includes only the portion of RMR payments that exceeds the value of energy produced from RMR units at the balancing energy price.

Figure 18: Expenses for Out-of-Merit Commitment and Dispatch 2003 and 2004



The figure shows that the sum of OOME costs and incremental energy costs from RMR units declined from $107 million in the first eight months of 2003 to $78 million in the first eight months of 2004, a decrease of 27 percent. The reduction in OOME costs primarily resulted from a 61 percent reduction in OOME Up costs, especially during the summer months of May through August. OOME Down costs decreased approximately 27 percent during the study period in 2004 from the comparable period in 2003.

The sum of out-of-merit capacity costs also decreased during the first 8 months of 2004. The total costs of OOMC and RMR units decreased from $175 million in 2003 to $130 million in 2004, a decrease of 26 percent. OOMC costs were lower throughout the study period, falling by approximately 39 percent in 2004. The RMR costs increased by approximately five percent in 2004 over the comparable period in 2003.

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

Although ERCOT allocates these costs on an ERCOT-wide load ratio share basis, they are incurred at specific locations on the network. The next two figures show the out-of-merit capacity and dispatch costs separately by location. Figure 19 summarizes the out-of-merit capacity costs and RMR capacity costs by location for 2003 and 2004. This figure shows RMR costs in the darker color and OOMC costs in the lighter color for non-RMR units. The striped areas in the chart are the OOMC costs paid to units that later became RMR units.

Figure 19: Expenses for OOMC and RMR by Region 2003-2004

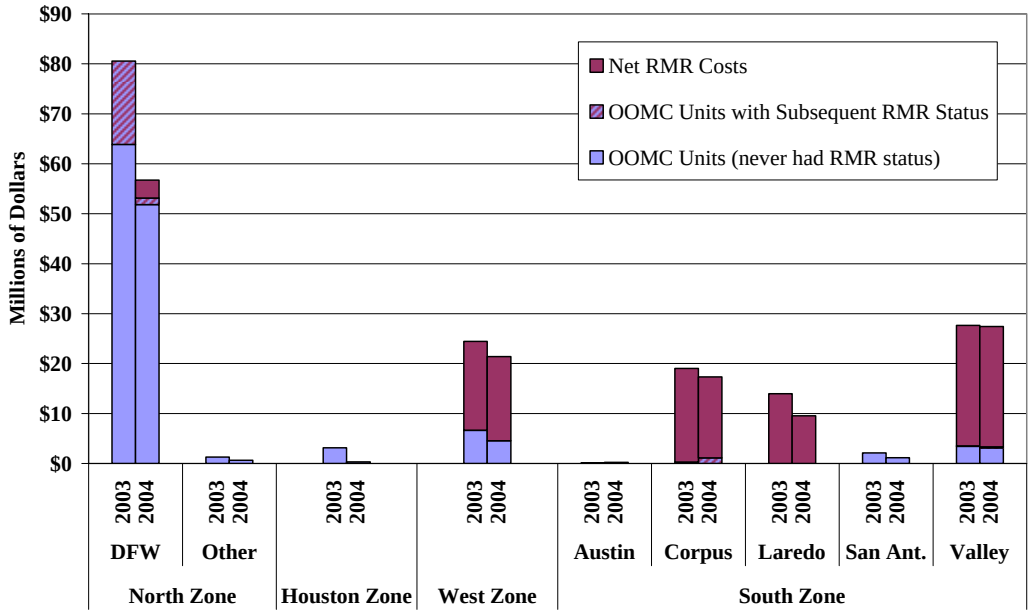


Figure 19 indicates that uplift costs for capacity changed very little outside the Dallas/Ft. Worth area from the first eight months of 2003 to the comparable period in 2004. The most significant change in the DFW area is that Eagle Mountain units 2 & 3 became RMR resources beginning May 2004.[13] After this point, these units received $3.6 million in RMR payments through August. The same two units accounted for approximately $16 million of the OOMC costs during the first eight months of 2003 and very little OOMC costs in early 2004, which the striped portions of the bars in the DFW area show. Figure 19 also reveals reductions in payments to other OOMC units in the Dallas/Ft. Worth and in other areas from 2003 and 2004.

According to the ERCOT Protocols, ERCOT pays OOMC units for (a) starting-up and (b) staying on-line. In February 2004, the compensation formulas changed for both components of OOMC payments. Previously, payment formulas for starting-up and staying on-line were not dependent on the balancing energy price, which caused problems for two reasons. First, there was no guarantee that the sum of the start-up payment, operating payment, and revenue from the balancing energy market would be sufficient for a unit to recover its costs.

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

The next analysis reviews the costs ERCOT incurred to dispatch generating resources out of merit to resolve local congestion. ERCOT incurs these costs in the form of OOME Up

[13] Eagle Mountain unit 2 only went on RMR status for a single month when Eagle Mountain 3 went off-line for maintenance.

and OOME Down payments, as well as payments to RMR resources for incremental energy above minimum generation. Figure 20 shows uplift costs for units by region for the first eight months of 2003 compared to the same months during 2004.

Figure 20: Expenses of Out-of-Merit Dispatch by Region 2003-2004

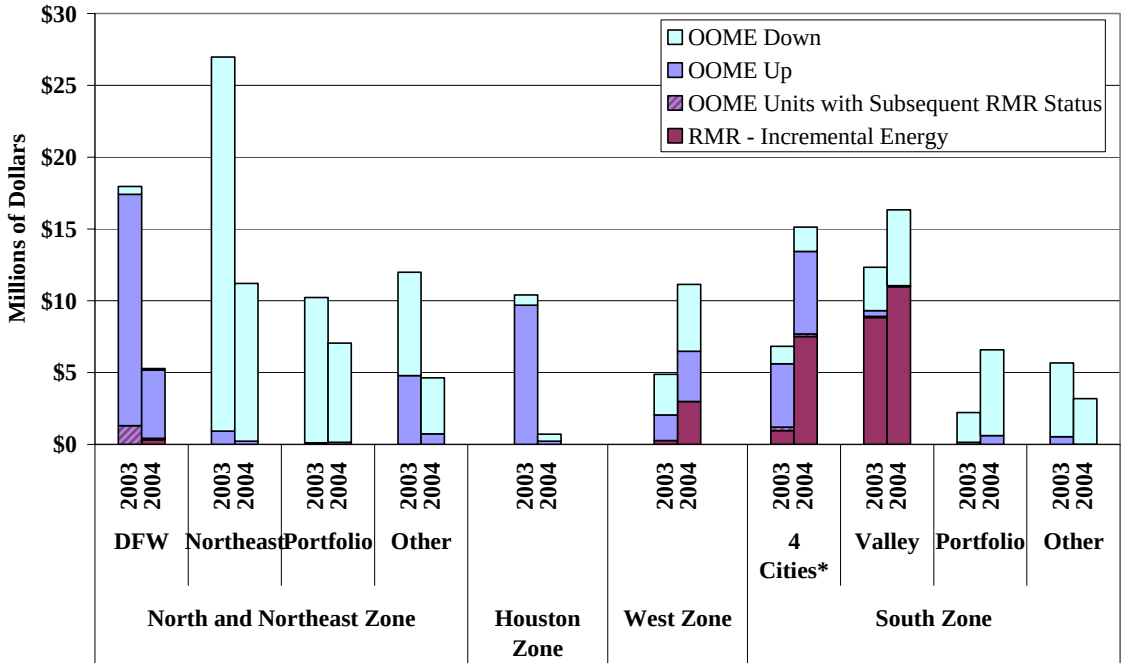


Figure 20 shows that uplift for OOME in the first 8 months of 2004 decreased significantly in the North, Northeast, and Houston zones, and increased substantially in the West and South zones relative to 2003. In the North, Northeast, and Houston, uplift for OOME Up deployments fell from $33 million to $6 million and uplift for OOME Down deployments fell from $45 million to $22 million.

The uplift payments for out-of-merit dispatch to resources in the Houston zone decreased 93 percent from the first eight months of 2003 to the comparable period in 2004. The dramatic reduction in out-of-merit dispatch is related to the creation of a new CSC in 2004. The North-to-Houston CSC allows the balancing energy market to resolve the congestion on the 345 kV lines directly connecting the North zone to Houston.

In the West zone and South zone, the portion of uplift paid to RMR units for incremental energy rose from $10 million to $21 million. In addition, uplift for OOME Up

deployments increased from $8 million to $10 million and uplift for OOME Down deployments increased from $14 million to $21 million.

In West Zone, the increase in OOME Up costs of $1.7 million was paid primarily to resources in the western part of the zone. OOME Down costs increased $1.9 million, in spite of a reduction in uplift for wind resources. The estimated costs for incremental energy from RMR units increased from less than $1 million to $3 million due to increased production from RMR units. In the first eight months of 2003, RMR units in the West zone produced 349 GWh while the same units produced 480 GWh in 2004.

In the South zone, uplift costs increased substantially in 2004 due primarily to increased costs for incremental energy from RMR units. The incremental energy costs for RMR units increased from approximately $1 million to $7 million within the most populous areas in the South zone and from $9 million to $11 million in the Rio Grande Valley.

To help understand the reduction in costs in the North and Northeast zones, Figure 20 shows the uplift costs separately associated with four categories of deployments: (a) deployments of resources in Dallas/Fort Worth, (b) deployments of resources in the area which eventually became the Northeast Zone in 2004, (c) portfolio deployments from the North and Northeast zones, and (d) unit-specific deployments from all other areas in the North Zone. Dallas/Fort Worth is a load pocket where ERCOT dispatches units up to relieve congestion. The Northeast is a generation pocket where ERCOT dispatches units down to relieve congestion. Before 2004, these areas were within the same pricing region and so congestion between the areas was managed as local congestion. The addition of the Northeast Zone has contributed to the reduction in OOME Down dispatch within the Northeast and OOME Up in DFW and other areas within the North zone.

We attribute some of the reduction in local congestion costs in 2004 to the suspension of the “Market Solution” method used to solve local transmission constraints. Prior to July 18, 2003, a Market Solution would exist whenever three or more unaffiliated suppliers were capable of relieving a local constraint. When a Market Solution existed, SPD would select the most cost effective resource(s) based on the shift factors and offer premiums for each resource. Resources that ERCOT incremented were paid the energy clearing

price plus their offer premium. Likewise during this period, ERCOT paid decremented resources their offer premium as compensation for the lost opportunity of producing at the market price.

ERCOT discontinued the Market Solution process because, in practice, the outcomes often were non-competitive and ERCOT frequently dispatched resources with offer premiums approaching $1,000/MWh. Market Solutions accounted for approximately $22.8 million in uplift payments from January to July 2003 to relieve relatively small amounts of congestion in real-time. While use of the Market Solution only occurred in 7.5 percent of intervals, they accounted for approximately 30 percent of the uplift for dispatch to manage local constraints during this period. The top five recipients of uplift for local balancing energy in 2003 are shown in Table 4.

Table 4: Comparison of Costs from Market Solutions vs. Non-Market Solutions Top 5 Recipients of Uplift for Market Solutions During 2003

	2003 Local Balancing Energy		2004 Out-of-Merit Energy	
Plants	**Uplift Payments ($ in millions)**	**Avg. Deployed Premium ($/MWh)**	**Uplift Payments ($ in millions)**	**Avg. OOME Payment ($/MWh)**
Kiamichi	$4.6	$996	N/A	N/A
Tradinghouse	$2.6	$173	$0.2	$4
Lake Hubbard	$2.3	$925	$0.5	$38
North Lake	$1.8	$462	$1.3	$35
Freestone Energy	$1.5	$150	$0.2	$17

Table 4 shows that the top five recipients of uplift from Market Solutions from January through July 18, 2003. SPD deployed these plants because the units were the most economic choice available, in spite of their extremely large offer premiums. For instance, when Kiamichi came online and created a generation pocket, ERCOT deployed the Kiamichi plant at an average premium of $996/MWh and accumulated $32 million in credits to its owners under the Market Solution.

However, the owners of Kiamichi and several others agreed to accept a much smaller payment (averaging about $150/MWh in the case of Kiamichi).[14] ERCOT developed a

[14] Public Utility Commission of Texas Memorandum, August 21, 2003, Project No. 25937, PUC Investigation into Possible Market Manipulation of the ERCOT Market.

more permanent solution to the problem in 2004 when it created the Northeast zone and a new CSC between the Northeast and North zones to price and directly assign the congestion costs to the participants using that transmission interface. The table shows the total payment and average payment per megawatt-hour for out-of-merit dispatch during the period from January to August 2004.

The 2004 OOME payments provide a benchmark for comparing the 2003 uplift that reflects the Market Solution method. For each of the five plants, the 2004 total payments are considerably smaller than the 2003 uplift from Market Solutions. Furthermore, the five plants were deployed in 2003 in spite of average offer premiums ranging from $150/MWh to $1000/MWh. In 2004, the same plants received $38/MWh or less for out-of-merit dispatch.

Under the Market Solution method, resources deployed for local constraints were paid their offer premiums and the resulting costs were not directly assigned to the market participants. This gave some market participants an incentive to raise their offer prices. In a well functioning market, competing offers exert discipline on market participants who would otherwise raise their offer prices.

Market participants can also have an incentive to "create" local congestion by submitting resource plans that show a level of output systematically different from its actual real-time output. For instance, a resource that frequently receives OOME Up instruction to manage local congestion may have an incentive to deliberately understate its planned generation to increase the probability of receiving an OOME Up instruction. Likewise, a resource that frequently receives OOME Down instructions to relieve local congestion may have an incentive to overstate its planned generation to increase its probability of receiving an OOME Down instruction. Thus, the incentive problems associated with the Market Solution and the fact the local congestion costs are not directly assigned can increase the frequency of local congestion, as well as raise local congestion costs per megawatt-hour. ERCOT has mitigated, but not eliminated the problem, by suspending the Market Solution and utilizing the formula-based OOME payments.

Lastly, the local congestion costs decreased in 2004 due to a change in the local congestion management process. Prior to June 2004, when individual units in a zone were dispatched up or down in a zone, other specific units were instructed to alter their output to counter-balance these deployments. Although the second group of generators was only redispatched to keep supply and demand in balance, they were paid premiums in the same manner as the first group of generators. In June 2004, the market was modified to deploy portfolio offers to counterbalance the local deployments. These portfolio deployments are compensated through the balancing energy prices and, thus, do not contribute to uplift.

In summary, there have been significant reductions in expenses for out-of-merit commitment and dispatch actions in 2004. These reductions are due to:

- The creation of a Northeast zone and the North-to-Houston CSC;
- The suspension of the Market Solution test; and
- The elimination of resource-specific deployments to maintain the energy balance in the local congestion process.

The formation of the new zone and CSCs has directly assigned congestion rents on these CSCs, shifting the burden of relieving congestion on these lines to the participants using these CSCs rather than uplifting the costs to all load in ERCOT.

Although these changes have reduced the local congestion costs, these costs for Dallas/Ft. Worth area continue to be very high, exceeded $60 million during the first eight months of 2004. ERCOT and the Congestion Management Working Group should continue to consider the potential benefits of creating a zone for Dallas/Ft. Worth.

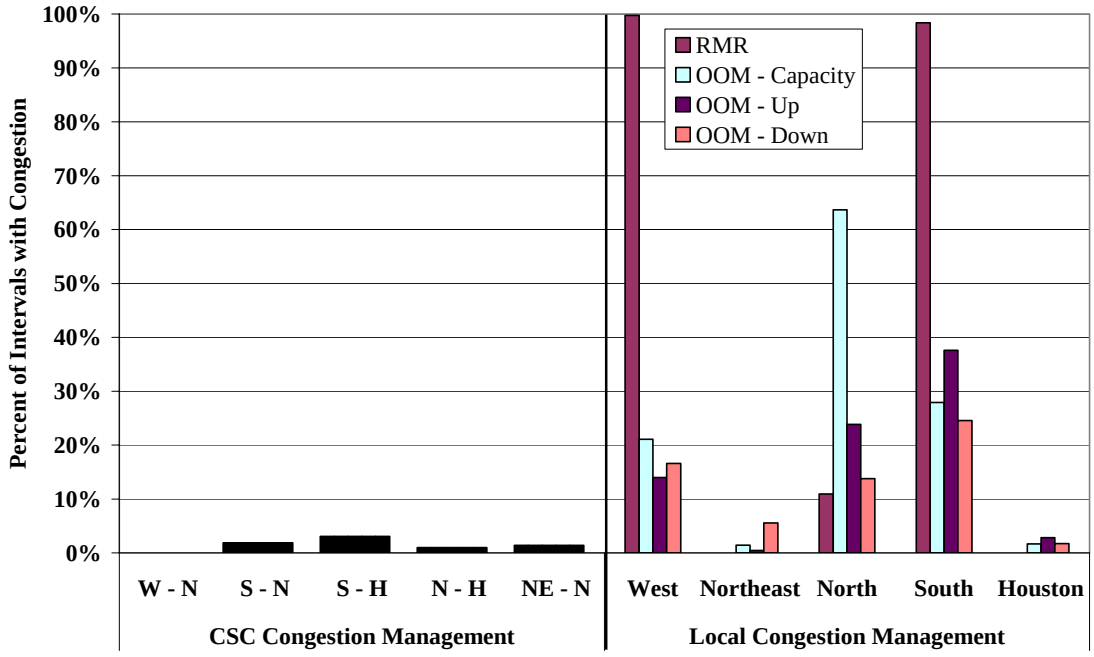
B. Comparison of Local and Interzonal Congestion

The zonal market is designed to manage interzonal congestion using a simplified network model with a small number of transmission interfaces. The intent in any zonal market should be to maximize the portion of the congestion that is managed by the zonal markets. Hence, most of the congestion in ERCOT's case should occur on the CSCs. Any remaining congestion management is performed through resource-specific

deployments. This sub-section of the report compares the relative amounts of interzonal and local congestion management from January to August 2004.

There is no straightforward way to compare the quantities of local and interzonal congestion costs because they occur in very different ways. In the case of local congestion, ERCOT redispatches and commits specific resources in order to lower flows over constrained lines to acceptable levels. In the case of interzonal congestion, the balancing energy market model adjusts prices and issues portfolio energy deployments in each zone until it finds a solution that respects the transmission limits on the CSCs. Given that there is no single way to compare the quantities of local and interzonal congestion management, this sub-section makes several different comparisons. Figure 21 shows the frequency of interzonal congestion management by CSC and local congestion management by zone from January to August 2004.

Figure 21: Frequency of Interzonal and Local Congestion
January to August 2004



The left panel in Figure 21 is a bar chart denoting the frequency of congestion on each CSC. The figure shows that interzonal congestion was relatively infrequent for each of the CSCs. The most frequently congested of these was the South-to-Houston interface, which was congested slightly more than 3 percent of the time.

The right panel of Figure 21 shows four side-by-side bars for each zone. The first bar for each zone shows the fraction of intervals where ERCOT deployed an RMR unit. The West and South zones had RMR units running in more than 98 percent of intervals, whereas the North zone had RMR units running in approximately 11 percent of intervals. The second bar shows how often ERCOT committed OOMC capacity. This occurred more than 20 percent of the time in the West zone, more than 62 percent of the time in the North zone, and approximately 28 percent of the time in the South zone.

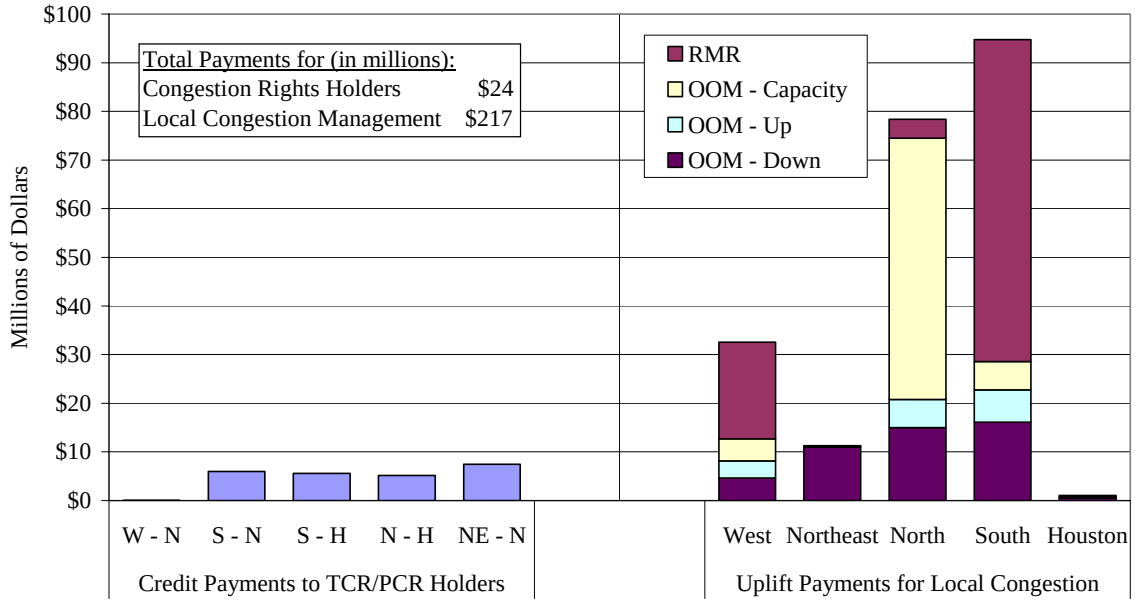
The final two bars in the figure show the frequencies of OOME Up and OOME Down deployments. They occur with the greatest frequency in the South zone where OOME Up occurred in 38 percent of the intervals and OOME Down occurred in 24 percent of the intervals. The frequency of OOME Up deployments for local constraints was 23 percent in the North zone, 14 percent in the West zone, less than 1 percent in the Northeast zone, and 3 percent in Houston. The frequency of OOME Down deployments for local constraints was 13 percent in the North zone, 17 percent in the West zone, 6 percent in the Northeast zone, and 2 percent in Houston.

Figure 21 shows that local congestion is far more frequent than interzonal congestion. One important reason that local congestion is far more frequent than interzonal congestion is that the costs of interzonal congestion are directly assigned to those using the CSCs while the local congestion is socialized. Without direct assignment, suppliers contributing to local congestion have no incentive to limit the flows they cause over the constrained transmission facility. Frequency of congestion, however, may not measure the overall economic significance of the congestion. In order to observe the significance of each type of congestion, Figure 22 compares the costs of interzonal and local congestion.

When a CSC constraint binds in real-time, the zonal prices will differ to reflect the economic value of the constraint. A congestion right holder is entitled to receive the economic value of the congestion on the CSC (i.e., the shadow price) times the quantity of TCRs held. For instance, the holder of 10 MW would receive $400 if the shadow price of the CSC between zones was $40/MWh for one hour. Congestion rights

payments are a useful way to measure the interzonal congestion costs. Hence, Figure 22 compares payments to CSC congestion rights holders of uplift payments to resources providing local congestion relief from January to August 2004.

Figure 22: Payments to TCR Holders vs. Local Congestion Payments January to August 2004



The left panel in Figure 22 shows total payments to TCR holders and PCR holders since the beginning of 2004. There have been virtually no payments to holders of congestion rights for the West-to-North CSC. However, credit payments for each of the other CSCs ranged from $5 million to $7.5 million through August 2004. Although congestion was substantially less frequent on the North-to-Houston CSC than the other CSCs, it has more transfer capability and a higher quantity of TCRs to satisfy when it is congested.

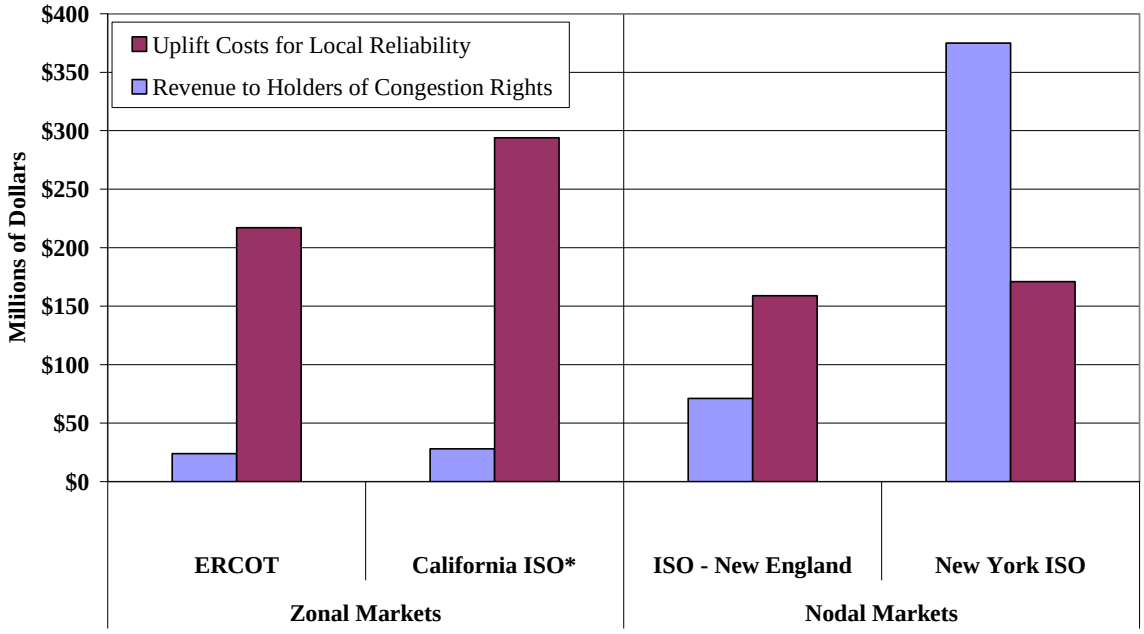
The right panel in Figure 22 shows uplift payments for each type of local congestion management in each zone. The South zone required the most uplift for local congestion, totaling $93 million. Most of these costs were paid to RMR resources. The North zone was the second highest-cost zone with $79 million in local congestion costs. In this zone, the largest payments were to units committed out-of-merit (OOMC). ERCOT paid units that relieved local congestion in the West zone $32 million, which is significant given the quantity of load and resources in the zone. Uplift for OOME ranged from $10 million to $22 million in each zone except for Houston.

In summary, the total payments for interzonal congestion were $24 million through August 2004. Despite reductions in local congestion costs from 2003, the local congestion costs were almost an order of magnitude higher than the interzonal congestion costs at $217 million. While these two quantities are not directly comparable, they do provide an indication of the relative significance of interzonal and local congestion.

Figure 21 and Figure 22 indicate that most congestion costs are associated with managing local rather than interzonal transmission constraints. This is due in part to the small number of zones and CSCs that ERCOT uses to manage congestion. Nodal markets are intended to fully reflect the costs of all constraints and local reliability requirements in market clearing prices. Even in a nodal market, however, there can be significant uplift costs for local reliability when the markets do not fully reflect all reliability requirements.

The following analysis compares the costs of market-based and out-of-market congestion management in nodal and zonal markets. Figure 23 shows market-based and out-of-market congestion management costs for ERCOT, the California ISO, the ISO – New England, and the New York ISO.

Figure 23: Comparison of Market-Based and Local Congestion Costs
Nodal Markets vs. Zonal Markets – Jan to Aug 2004



* For the CAISO, the bars show 2003 annual figures and congestion rents are shown rather than payments to congestion rights holders.

Comparing the total costs of congestion management between markets can be misleading because these costs depend on many factors that are unique to a particular market. Figure 23 indicates, however, that uplift costs for local reliability are significantly higher in the zonal markets than in the nodal markets. In ERCOT and the California ISO, local congestion costs were between nine and ten times larger than the market-based congestion costs. In the New York ISO, local reliability costs were half the size of market-based costs.

As expected, out-of-market congestion costs are far more significant in zonal markets than nodal markets. The size of out-of-market congestion costs, however, is still quite substantial in nodal markets. In the New England market, the vast majority of uplift is incurred to maintain local reserves requirements. The New England market runs day-ahead and real-time nodal energy markets, but does not jointly optimize the procurement of energy and operating reserves. Thus, the prices in the New England market reflect the value of energy, but not the operating reserves needed in specific areas.

Likewise, the New York ISO must satisfy local reliability requirements in New York City. Although the New York ISO does have jointly-optimized energy and ancillary services markets, the local requirements for NYC are not defined in these markets and must, therefore, be satisfied by out-of-market actions. In sum, nodal markets can minimize uplift by using jointly-optimized energy and reserves requirements that fully reflect local transmission and reliability requirements. However, ERCOT can significantly reduce the uplift costs in its current zonal market design by expanding its set of zones and CSCs.

The analysis in this subsection of the report shows that the majority of transmission congestion in ERCOT does not occur on CSCs. Local congestion that is not resolved using CSCs is not directly reflected in zonal clearing prices. Hence, one may conclude that most of the economic value of congestion in ERCOT is not reflected in the zonal balancing energy prices. This is a significant concern because it indicates that the current market prices are not efficiently and transparently revealing the direct and indirect costs of congestion in ERCOT. This result of the zonal market design has short-term effects on

production and consumption, as well as long-term effects on investment and retirement decisions. The fact that most congestion costs are recovered through uplift charges that are socialized across the ERCOT region means that:

- ERCOT does not compensate resources that are important in relieving network constraints consistent with their value to the system, which will limit investment that would otherwise occur in congested areas;
- ERCOT effectively over-compensates resources that contribute to local congestion because the zonal price they receive does not reflect the economic value of the loadings they place on the local constraints;
- Loads that have an ability to respond to price signals will not receive accurate economic signals relating to their effects on local transmission constraints; and
- Loads and other market participants have a limited ability to hedge the costs of congestion they face.

Under the current market design, these issues can only be addressed by improving the definition of the zones and CSCs in ERCOT. To that end, we have recommended a number of changes to the process in ERCOT for defining zones and CSCs in the Interzonal Congestion Section above.

In the longer-term, the implementation of nodal electricity markets would most comprehensively resolve each of these issues because the prices at each location would accurately reflect both the local and interzonal constraints in the current system. In addition to improving the accuracy of the price signals, the nodal markets would allow the loads to fully hedge congestion costs by holding transmission rights and contracting bilaterally for delivery of power at specific locations.

In the next subsection, we critically evaluate the congestion management processes in ERCOT that attempt to separate the zonal deployment of supply to manage interzonal congestion from the redispatch of specific resources to resolve local congestion. The interaction of these processes can have significant effects on ERCOT's market outcomes.

C. Multi-Step Congestion Management Process

To resolve local congestion, ERCOT solves the balancing energy market in three steps: (i) the first step determines the dispatch levels based on zonal portfolio schedules and offers to meet demand while observing interzonal transmission limits, (ii) the second step

makes incremental/decremental dispatch changes to specific resources to manage local congestion, and (iii) the third step uses portfolio offers to counter-balance changes from the second step. This is done by re-clearing the balancing energy market while considering the interzonal and active local constraints from Step 2. We describe these three steps in detail below.

Step 1

Step 1 determines the balancing energy deployments needed to meet the energy demand and resolve any interzonal congestion. In this step, ERCOT ignores local congestion and only considers the flows over the CSC interfaces. The energy deployments in each zone are based on portfolio offers and not associated with individual resources. If ERCOT detects no local congestion, ERCOT finalizes QSE portfolio deployments.

The only local constraints SPD evaluates are those that have been "activated", which occurs when a constraint becomes binding and the units that have non-trivial impacts on the flow over the constrained facilities are entered into SPD. ERCOT enters a GSF for each generator that has a non-trivial effect on the constraint. An operator can manually activate a constraint, or allow the SFT software to activate the constraints.[15]

To evaluate whether a local constraint is binding, ERCOT must allocate the portfolio balancing energy deployments to individual units (i.e., portfolio decomposition). ERCOT makes this allocation to the units on a pro rata basis according to the available undispatched output (or "head room") on each of the online resources in the portfolio. This allocation of the balancing energy deployments augments the resource plan information the QSEs provide to estimate the output level for each unit.[16] These output

[15] ERCOT has been running the SFT software on a trial basis since July 2004.

[16] ERCOT implemented a Constraint Oscillation Management mechanism in June 2004 that allows the SPD model to incorporates previous interval resource instructions when clearing for the next interval. Estimate of a unit initial starting point for the next interval is based on the previous interval's deployment and resource plan. This improvement results in more accurate portfolio decomposition, and significant reduction in instances of constraint oscillation. Constraint oscillation had occurred when SPD generated resource-specific instructions one interval to resolve a local constraint that were not given in the next interval because the constraint was no longer binding, causing the constraint to bind once again.

levels allow SPD to assess whether the local constraint will be binding in the next interval. If any local congestion is present, ERCOT proceeds to Step 2.

Step 2

Step 2 uses unit-specific offer premiums to select units to increase output (increment) and decrease output (decrement) in order to reduce the flow on the constrained facility and manage the local congestion. The increase and decrease in output is assumed to occur from the estimated unit loadings that result from the portfolio decomposition we describe in Step 1. In the past, ERCOT had defined the offer premiums as premiums above the prevailing balancing energy price, but it has redefined the offer premium as the entire incremental/decremental offer price for the resource. ERCOT currently determines the offer premiums based on standardized formulas that are technology-specific.

Using these unit-specific offers, the model redispatches "effective" generators (those with a non-trivial GSF for the constraint) and "ineffective" generators (all other generators). The model's objective is to minimize the total cost of the redispatch, defined as the incremental output times the incremental offer premiums *less* the decremental output times the decremental premiums. The model also maintains the energy balance and respects the interzonal constraints. Although the model may select both effective and ineffective generators to redispatch, only the effective generators are constrained to provide local congestion relief in Step 3 (the instructions for the ineffective generators are discarded). In other words, in Step 3, ERCOT generally cannot (a) decrement effective generators that receive incremental instructions in Step 2, or (b) increment effective generators that receive decremental instructions in Step 2.

Step 3

Step 3 re-clears the balancing energy market using the QSEs' portfolio offers, respecting the instructed deviations derived in Step 2, the power balance constraint, and the interzonal constraints. This step will establish zonal energy prices that can differ substantially from the prices calculated in Step 1 when large amounts of energy for

counter-balancing is needed. Step 3 also establishes the final balancing energy deployments for the interval.

These three steps were developed presumably to preserve the integrity of the zonal market by separating the actions taken to manage local congestion from the actions taken to manage interzonal congestion. As we will show below, however, the redispatch for local congestion can profoundly affect the balancing energy market. Therefore, it should be done as efficiently as possible, recognizing the interactions between the local congestion management process and the zonal balancing energy market. To that end, we propose a significant change in this multi-step process below. First, however, we assess the effects that local congestion management actions can have on the balancing energy market under the current process.

D. Impact of Local Congestion on Balancing Prices

In an electricity network, all elements are inter-related, and so management of local congestion will inevitably impact the balancing energy market. Therefore, it is not efficient to manage local congestion in isolation without considering its effects on the balancing energy market and interzonal congestion. Failing to *simultaneously* consider all factors leads to higher production costs than are necessary and can significantly alter the zonal balancing energy prices

For example, suppose that ERCOT can manage a local constraint from point A to point B either (i) by increasing output from a resource at point A at a cost of $75/MWh or (ii) by decreasing output from a resource at point B at a savings of $20/MWh. For each megawatt increase from the resource at point A, ERCOT must reduce balancing deployments by one megawatt. Likewise, for each megawatt decrease from the resource at point B, ERCOT increase balancing energy deployments by one megawatt.

If the marginal offer price is $60/MWh, the marginal cost of increasing output at point A is $15/MWh (= $75 cost for energy at point A minus $60 saved from a portfolio offer). Similarly, the marginal cost of decreasing output at point B is $40/MWh (= $60 cost for portfolio energy minus $20 saved from decreasing output at point B). Based on the price of the marginal portfolio offer, the cost of incrementing the unit at point A is much less

expensive than decrementing the unit at point B. Thus, without knowing the price of the marginal portfolio offer, ERCOT cannot know whether incrementing at point A or decrementing at point B is more efficient. The current design of the balancing energy market redispatches units to resolve local congestion without considering the cost of the marginal portfolio offer.

Even when done optimally, local congestion management can substantially effect market outcomes. For every megawatt dispatched down for local congestion, ERCOT must deploy an additional megawatt from a portfolio offer. If ERCOT decrements 1,000 MW to resolve a local constraint, it must deploy an additional 1,000 MW from portfolio offers, which is likely to increase the market clearing price. For every megawatt ERCOT increments to resolve local congestion, ERCOT needs to deploy one fewer megawatt from portfolio offers, which may depress the market clearing price.

The following analysis demonstrates the impact of local congestion management on market clearing prices. We selected the 133 intervals from June through September 2004 when ERCOT needed to resolve local congestion at a time when the zone-weighted average balancing energy price was $80/MWh or more. For these intervals, we re-solved the balancing energy market without regard for local congestion (i.e., we turned off local constraints and treated resources that ERCOT would have flagged for manual OOME as if they were not flagged). Through this process, we estimate what market clearing prices would have been if there were no local congestion. Table 5 summarizes the results of this analysis.

Table 5: Effects of Local Deployments on Balancing Energy Prices

Original Price Range	Direction of Local Deployment	Number of Intervals	Avg. Net Local Deployment	Avg. Load	(1) Avg. Price	(2) Avg. Price without Local	= (2) - (1) Avg. Difference
> $150	Up	6	611	49,709	$202	$220	$18
	Down	49	-1308	46,564	$210	$79	$131
$100 to $150	Up	13	314	49,101	$129	$147	$18
	Down	24	-1139	45,672	$128	$66	-$62
$80 to $100	Up	5	188	50,283	$83	$85	$1
	Down	36	-965	47,336	$87	$70	$18

We divided the intervals studied in this analysis into three price ranges: $150/MWh and above, $100/MWh to $150/MWh, and $80/MWh to $100/MWh. Each of these groups is divided into "Up" or "Down" categories based on the prevailing direction of deployments for local congestion. Table 5 shows a total of 109 intervals with net down deployments and only 24 intervals with net up deployments. The average magnitude of net deployments is significantly greater for the down direction. This is not surprising because frequently only small amounts of unloaded capacity exist in load pockets to increment while ERCOT has large quantities of decremental energy available in other areas.

Table 5 also shows that the average load is higher in the "Up" intervals than in the "Down" intervals. Because local down deployments make significant amounts of capacity unavailable that would otherwise be economic, balancing energy prices can be very high in these intervals even though the load is lower.

Table 5 reports the average balancing energy price, as well as the balancing energy price that was determined by running the balancing energy market model without local congestion management.[17] The difference reported in the final column is the average estimated price impact from assuming no local congestion. In intervals with upward deployments for local congestion, the clearing price is lower than if there was no local congestion. Conversely, in intervals with down deployments for managing local congestion, the clearing price is much higher than if there was no local congestion. This is because local down deployments effectively remove inexpensive energy from the balancing energy market, causing ERCOT to strike more expensive portfolio offers.

Removing local congestion management from the "Down" intervals originally priced above $150/MWh would cause the average balancing energy price to decrease from $210/MWh to $79/MWh. In "Down" intervals originally priced from $100/MWh to $150/MWh, removing local congestion decreases the average balancing energy price by

[17] The balancing energy prices reported are after price corrections by the dispatch software but before price corrections performed by ERCOT. In cases where the balancing energy price was greater than $300/MWh, we used $300/MWh in the table summary.

more than half from $128/MWh to $66/MWh . The magnitude of the price impact of local congestion management is larger at higher price levels, because the supply is more price inelastic under tight conditions.

E. Local Congestion Conclusions

Although we performed the analysis in the prior subsection on a relatively small number of intervals, we can draw several conclusions from the results. First, local congestion management can have large indirect effects on portfolio deployment and the market clearing price. Decrementing resources for local reasons reduces the capacity available to the balancing energy market, tending to increase balancing energy prices; incrementing resources for local reasons increases output in the balancing energy market, tending to reduce balancing prices.

Second, the analysis suggests that the multi-step congestion management process that the balancing energy market model uses is vitally important for an efficient dispatch and accurate balancing energy prices. ERCOT can resolve a local constraint by reducing output on the generation-side or increasing output on the load-side of a constraint. However, when the model currently counter-balances local deployments with portfolio balancing energy deployments (in Step 3), it does not consider the marginal costs of such deployments. Hence, when load is high and portfolio offers are in short supply, the opportunity costs of decrementing low-cost units is extremely high since the output from these units must be replaced by very high-cost portfolio offers, which can lead to price spikes and market inefficiencies. ERCOT could avoid such inefficiencies by taking into consideration the marginal cost of the portfolio balancing energy deployments needed to counter-balance the OOME Down deployments and by deploying OOME Up resources instead when this would lower overall costs.

ERCOT recently recognized that it was using large quantities of OOME Down deployments to manage certain local congestion. ERCOT attempted to address this issue by raising the threshold for resources that it would redispatch to resolve local congestion from a 2 percent shift factor to 4 percent shift factor. However, this change did not address the underlying modeling problem. ERCOT was compelled to reverse this change

because it limited the model's access to resources needed for local congestion management.

Our recommendation to address these issues would be for ERCOT to modify the multi-step balancing energy market optimization to minimize the costs of deployments for both local congestion and interzonal congestion. The original logic underlying the multi-step process, that actions to relieve local congestion should not affect the balancing energy market, is fundamentally flawed. Hence, the current market should be improved by recognizing the significant interaction between local congestion management and the balancing energy market and making the necessary changes to SPD to make the entire process as efficient as possible. Although it may be possible to develop a single-step methodology to optimize the local and interzonal deployments, we believe that an iterative methodology to achieve the lowest cost solution and efficient zonal energy prices will likely be needed. We recommend that ERCOT pursue this change as soon as possible because it should significantly improve the efficiency of the pricing in the current balancing energy market.

These issues will not exist under the Texas nodal markets that are under consideration since all transmission constraints will be economically managed through the nodal market. However, such markets, even if approved, will not be implemented for a considerable amount of time. Hence, it is important to take steps to ensure the current markets operate as efficiently as possible.

Earlier in this section, we concluded that the magnitude of local congestion that have prevailed in ERCOT far exceed the magnitude of interzonal congestion. This raises fundamental questions regarding the definition of the zones and CSCs in ERCOT. We discussed this issue in Section II and would reiterate the recommendations in that section regarding the process for modifying the definition of zones and CSCs in ERCOT.

III. LOAD FORECASTING

A. Day-ahead forecasting

Prior to each operating day, ERCOT produces a load forecast for each congestion management zone. The forecast is based on the weather forecast and a variety of other factors. The forecast is publicly available to market participants to assist them with commitment and scheduling decisions. An accurate day-ahead forecast reduces uncertainty and helps market participants make efficient decisions that reduce the cost of serving load and providing ancillary services. This subsection of the report evaluates the accuracy of day-ahead load forecasting relative to the performance of other system operators.

The analysis in this section of the report uses two metrics to quantify the accuracy of forecasting. The first measure is the average forecasted load minus the average actual load ("forecast-actual"). This captures systematic differences between the forecast and the actual loads. Thus, a positive number denotes over-forecasting on average whereas a negative number denotes under-forecasting. The second measure is the average of the absolute value of the difference between the forecasted value and the actual value ("forecast error"). This measures the magnitude of error regardless of the direction. Therefore, if all differences between the forecast load and actual load are random, we would expect that the average forecast-actual metric to be close to zero, while the forecast error could be relatively high.

It is most important to accurately forecast load during the highest load hour. To avoid emergency actions in real-time, the market must commit sufficient capacity to serve load and provide reserves under peak load conditions. Many market participants base at least part of their commitment decisions on day-ahead forecast results, committing a larger share of their resources when the forecast is high. An accurate day-ahead forecast enables them to do a better job deciding which units are likely to be economic. Figure 24 reports the accuracy of the day-ahead load forecast relative to actual real-time load in the peak of the day.

Figure 24: Forecast Error at Daily Peak
January to August 2004

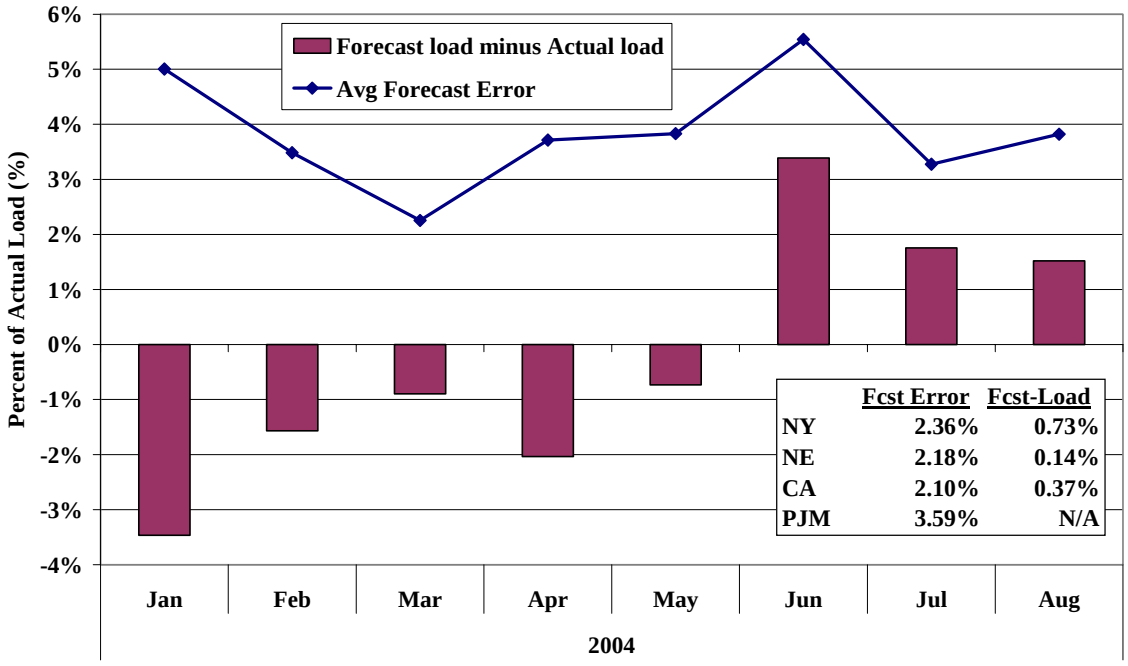


Figure 24 shows the two measures of the forecast accuracy in the day-ahead for each month from January to August 2004. The bars denote the systematic difference between the forecast and actual load. The forecast load was systematically lower than the actual load during the winter and spring, particularly January. In the summer months, the forecast load was systematically high by 1.5 to 3.3 percent. The line in the figure shows the average magnitude of average forecast error in each month. The average forecast error ranged from 2.2 to 5.6 percent over the 6 month period.

To provide a comparative evaluation of the ERCOT results relative to other regions, the table in Figure 24 shows summary information on day-ahead forecast accuracy for the New York ISO, the ISO–New England, the California ISO, and PJM ISO. The average forecast error was 3.9 percent in ERCOT, compared with a range of 2.10 percent to 3.59 percent in the other four regions. Hence, its forecast error was slightly higher than PJM and significantly higher than the other regions. This may be due in part to larger weather-related uncertainties in Texas than in the other regions.

The other metric, the average difference between the forecast and actual load, indicates that ERCOT's performance is in the same range as the other markets. This metric indicates that ERCOT over-forecasted load by 0.11 percent on average over the 8 month period. This is comparable to the California ISO and ISO–New England and significantly less than the New York ISO.

It is important to accurately forecast load during every hour of the day in order to have sufficient capacity on-line to serve load and provide reserves at all times. The timing of some commitments can be influenced by day-ahead forecast results by indicating when resources will be needed to serve load and when a unit may be economic. Figure 25 reports the accuracy of the day-ahead load forecast by time of day relative to actual real-time load.

Figure 25: Day-Ahead Load Forecast by Hour
January to August 2004

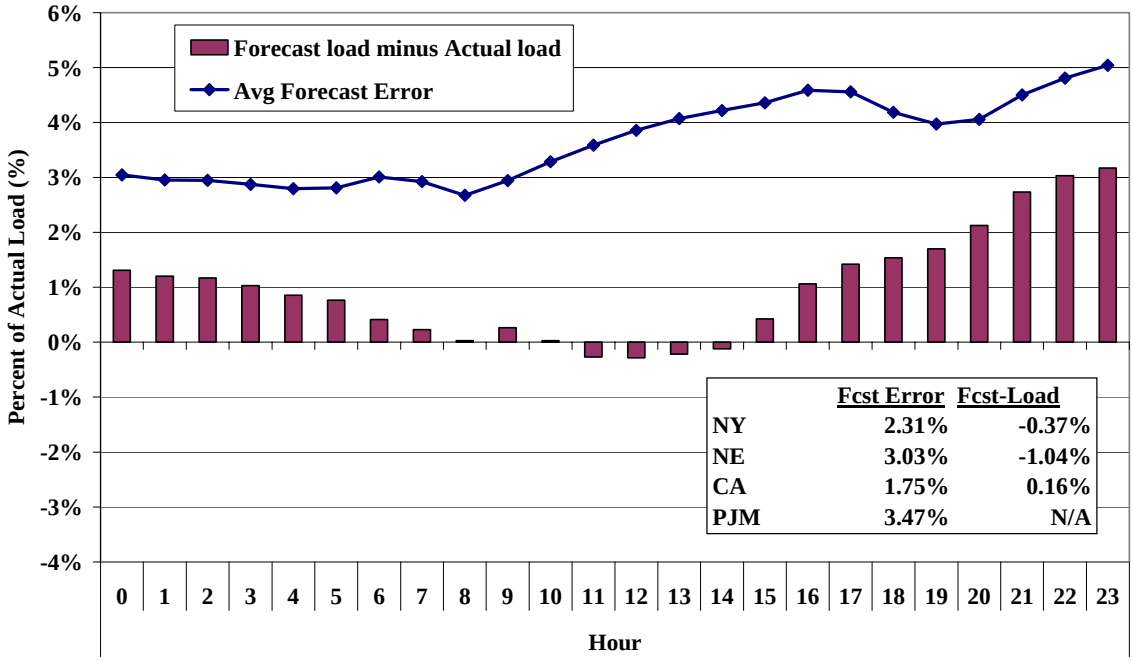


Figure 25 shows the same two measures of forecast accuracy as shown in prior figures, but shows them by hour of the day from January to August 2004. The bars denote the average difference between the load forecast and the actual load. The figure shows that the forecast load and actual load were comparable during the afternoon. In the late evening and early morning, the forecast was systematically higher by 1.1 to 3.1 percent.

Overall, there is a pattern that indicates that the forecast load is similar to the actual load during the high demand hours, although it does not decrease as low as the actual load overnight on average. Because there is a recognizable pattern to these differences, we recommend that ERCOT consider whether changes can be made to the forecast model that would limit this pattern.

The line in Figure 25 shows the average magnitude of forecast error in each hour of the day. The average forecast error ranged from 2.7 to 5.1 percent by the hour of day. The average error is generally higher in the higher-load hours during the afternoon, which is consistent with expectations because the uncertainty regarding the load in these hours is slightly higher due to weather and other factors.

To allow for comparisons to other markets, the table in Figure 25 shows summary information on day-ahead forecast accuracy in all hours for ERCOT, the New York ISO, the ISO–New England, the California ISO, and PJM ISO. These results are different that the tabular results shown in Figure 24 because that figure only included the forecast in the peak hour of the day, rather than all hours. The average forecast error was 3.74 percent in ERCOT, compared with a range of 1.75 percent to 3.47 percent in the other four regions. Like the results for the peak hour, the forecast error in ERCOT was higher than the error in other markets. However, the average difference between the forecast load and actual load was only 0.98 percent, which is comparable to the magnitude of systematic error for the ISO–New England although the ISO-New England systematically under-forecasted.

B. Real-time forecasting

The balancing energy market model (i.e., the SPD) schedules energy to meet a load quantity entered by the operator every 15 minutes. The final load quantity, SPD load, is the sum of the short-term load forecast ("STLF") and operator offset. The short-term load forecasting model updates the forecast approximately 25 minutes prior to the start of each balancing interval. The operator offset is based on three factors. First, the operator updates the Operating Period Desk ("OPD") forecast and to the extent this differs from the STLF, the difference is used to determine one part of the offset.

Second, if the sum of QSE energy schedules exceeds the planned generation in their resource plans by more than 500 MW, the difference is used to determine the second part of the offset. Third, operators review information on regulation deployment, frequency, schedule control error ("SCE"), time-of-day factors, changes in weather patterns, and other factors related to load to make any additional adjustments to the offset value.

A precise breakdown of the three parts of the offset was unavailable from ERCOT. Hence, we could not determine how the OPD forecast differed from the STLF. The net load used by SPD most closely reflects the OPD forecast load of any available data. Because the OPD forecast load itself and the data to calculate it are both unavailable, this sub-section analyzes deviations between the actual load and SPD load.

It is important to accurately forecast load to ensure that the balancing energy market model schedules sufficient energy to serve demand in each 15 minute interval. Having an accurate real-time forecast minimizes the need for regulation and reserves deployment. Figure 26 reports the deviation between SPD load and actual load by time of day.

Figure 26: SPD Load vs. Actual Load
January to September 2004

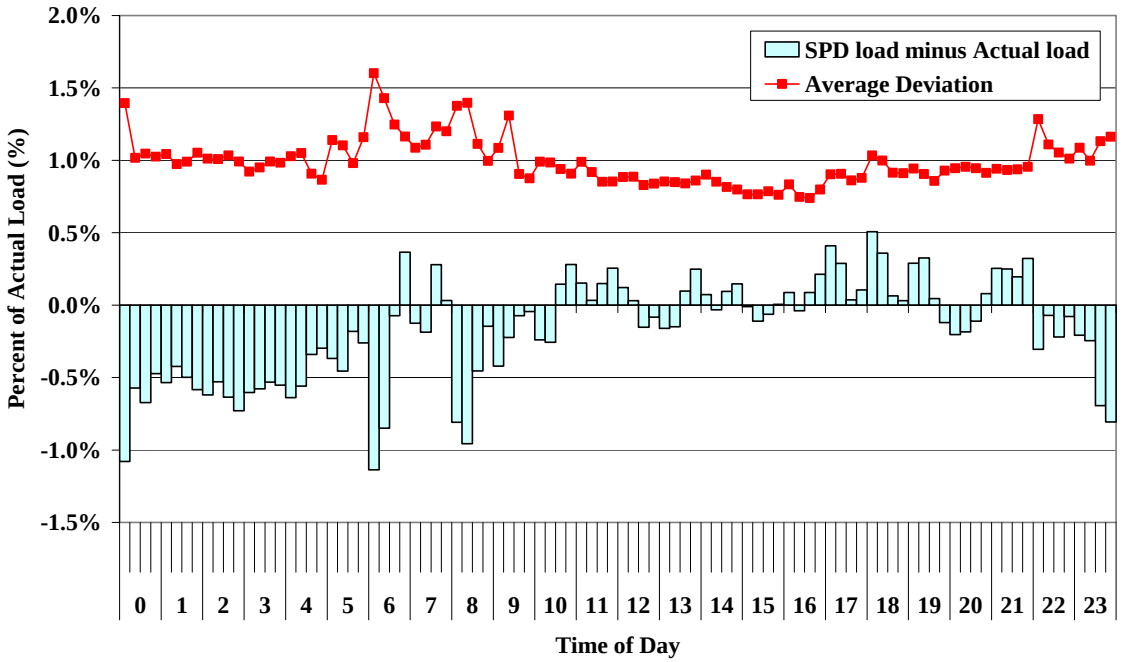
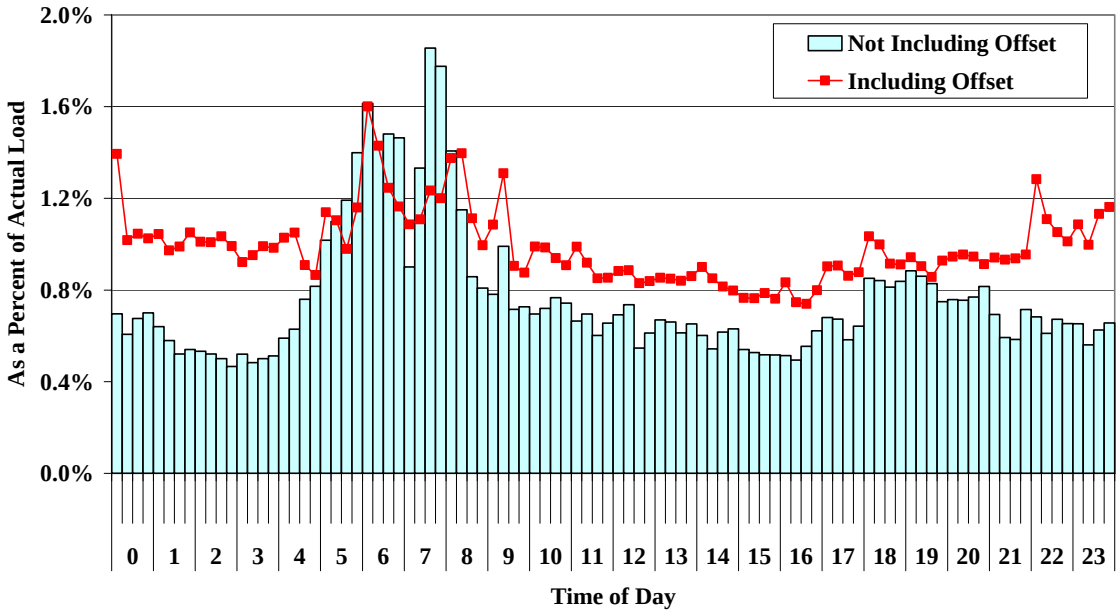


Figure 26 shows two metrics of deviations between SPD load and the actual real-time load for each interval of the day from January to August 2004. The bars denote the systematic difference between average SPD load and average actual load. The average SPD load is close to the average actual load during the afternoon. In the late evening and early morning, the SPD load was systematically lower than the actual load by 0.5 to 1.1 percent. The 30-minute periods from 6:00 am to 6:30 am and 8:00 am to 8:30 am show that the SPD load is approximately 1 percent lower than the average load.

The line shows the average magnitude of deviations between SPD load and actual load in each hour of the day (i.e., the average of the absolute value of the differences). The average load deviation was approximately 1 percent for most of the day. However, the average forecast error peaks at 6:00 am, 8:00 am, 10:00 pm, and midnight, rising to more than 1.5 percent. These deviations may be due in part to other adjustments made in the offset that do not relate to the short-term load forecast. For example, ERCOT operators may adjust the offset to account for systematic schedule control error by suppliers during certain intervals during the morning pick-up (e.g., 6:00 am). Even if these adjustments tend to improve the overall performance of the balancing energy market, they will tend to increase the deviations shown in the figure.

The operator offset is designed, in part, to improve upon the accuracy of the short-term load forecast. In the next analysis, we evaluate whether the offset improves the accuracy of the SPD-modeled load by comparing the accuracy of the short-term load forecast to the accuracy of SPD load that incorporates the operator offset. It is not straightforward to make this comparison because the offset includes adjustments for reasons unrelated to the load forecast. As noted above, many of these adjustments are intended to bias the load forecast in order to counter-balance other operational factors. While this improves system operation, it will tend to make the SPD load appear to be less accurate than it would be if the bias were removed. Figure 27 reports the forecast deviation with and without the offset adjustment. The deviation based on the SPD load includes the offset adjustment and the deviation based on the short-term load forecast does not.

Figure 27: Average Deviation With and Without Offset Adjustment
January to September 2004



The bars in Figure 27 indicate that without the offset, the average load deviation ranges from 0.5 percent to 0.8 percent for most of the day. However, the magnitude of the deviation starts to grow beginning at 4:00 am. The deviations peak in the interval that begins at 6:00 am at 1.6 percent and again in the intervals from 7:30 am to 8:00 am at 1.8 percent. It is reasonable to expect that the deviations will be higher in these ramping intervals since the uncertainty is higher when loads are increasing rapidly.

The line in Figure 27 indicates that with the offset, the average SPD load deviations are higher than the STLF deviations, ranging from 0.8 percent to 1.0 percent for most of the day. Hence, the offset causes the SPD load to less accurately reflect actual load. Like the STLF deviations, the SPD load deviations rise rapidly during the morning load pick-up, although less dramatically. Between 5:30 am to 8:15 am, the magnitude of SPD load deviation is generally less than the STLF deviation that excludes the offset. Hence, the offset improves the accuracy of the SPD load in these intervals.

Overall, SPD load appears to be less accurate than the short-term load forecast in the majority of hours, with the exception of the morning pick-up when it is slightly better. However, the analysis in the next section of this report suggests that the portion of the

offset attributable to factors other than changes in load forecasts is relatively significant. To the extent that this portion of the offset is accurately calculated and improves the performance of the balancing energy market, the fact that it may increase the SPD load deviation does not raise significant concerns.

We would note that a portion of the deviations evaluated above may actually be due to the error in forecasting the system loss factor. ERCOT defines load as the actual power consumed by end-users plus the losses that accrue from flows across the transmission and distribution systems. The system loss factor is a function of where load and generation are located. While the location of load is relatively consistent from day-to-day, the operator does not necessarily know which generators each QSE will run in order for the QSE to meet its schedule. During many portions of the day, the system loss factor stays relatively constant. However, during the morning load pick-up, it can be difficult to predict which resources will come on-line and how quickly. Thus, system loss forecast error likely gets bigger at the same time that the SPD load deviation is the largest. This issue is inherent to the current zonal market framework because deployments are made on a portfolio basis rather than a unit specific basis.

IV. REAL-TIME MARKET OPERATIONS

The fundamental requirement of the real-time operations is that supply continuously matches demand. Failing to do so can result in blackouts, damage to electrical equipment, and other problems. To accomplish this, the real-time market and ERCOT operators take the following steps:

- Prior to each 15 minute interval:
  - ERCOT estimates the load that it must serve (the short-term load forecast plus the offset), which we refer to as SPD load.
  - The SPD model compares the scheduled generation to this load and deploys the lowest cost balancing energy available to meet the SPD load, subject to the interzonal constraints and portfolio ramp constraints.
- During the 15 minute interval:
  - Because the actual load will vary during the interval and generators may not produce the expected level of electricity (SCE), ERCOT deploys regulation every four seconds to ensure that load matches generation.
  - However, regulation resources do not always respond reliably to the regulation signals and the signal itself can be inaccurate. This will create a residual error referred to as the Area Control Error ("ACE") that causes the frequency on the system to fluctuate.

If frequency fluctuates significantly enough, ERCOT can rely on responsive reserves to restore the supply demand imbalance. Responsive reserves include generation and Loads acting as Resources (LaaRs) with under-frequency relays ("UFRs") that trip when the frequency falls below a pre-determined threshold.

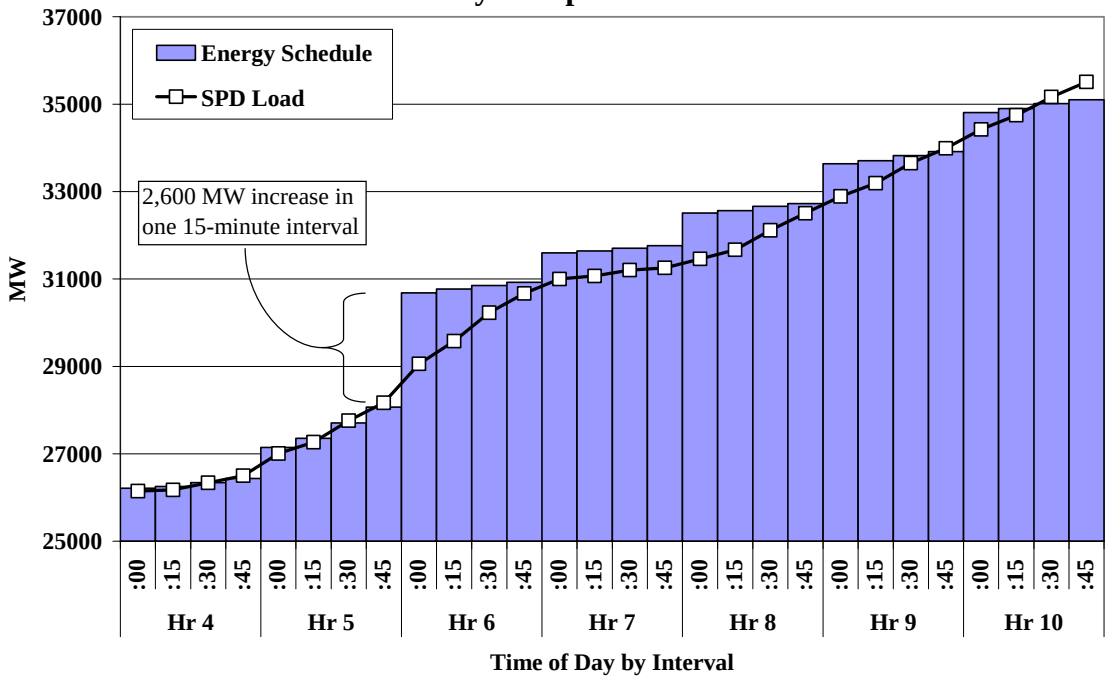
The prior sections evaluated the operation of the market related to how efficiently congestion is managed and how accurately ERCOT forecasts the load. This section examines the other steps in the real-time market and operations. The first subsection summarizes the performance of the balancing energy market, including an examination of scheduling patterns, balancing energy deployments, and balancing energy prices. This analysis is an update of similar analysis that we presented in the *2003 State of the Market Report for the ERCOT Wholesale Electricity Markets.*

The following subsections examine the deployment of regulation, generators' SCE, and the system ACE. Because we were able to obtain operating data for September 2004 in time to include in the analyses in this section, the period studied in this section extends through the end of September. Based on this analysis, we make recommendations intended to improve the performance of the real-time system. At the end of this section, we also examine QSEs' provision of ancillary services and the procedures for committing and dispatching peaking generators, such as gas turbines.

A. Scheduling and Balancing Energy Market Outcomes

We begin our analysis by examining factors that determine the demand for balancing energy during periods when load is ramping up and periods when it is ramping down. Figure 28 shows average energy schedules and actual load for each interval from 4:00 am to 11:00 am during 2004. In general, energy schedules that are less than the actual load result in balancing up energy deployments while energy schedules higher than actual load result in balancing down energy deployments.

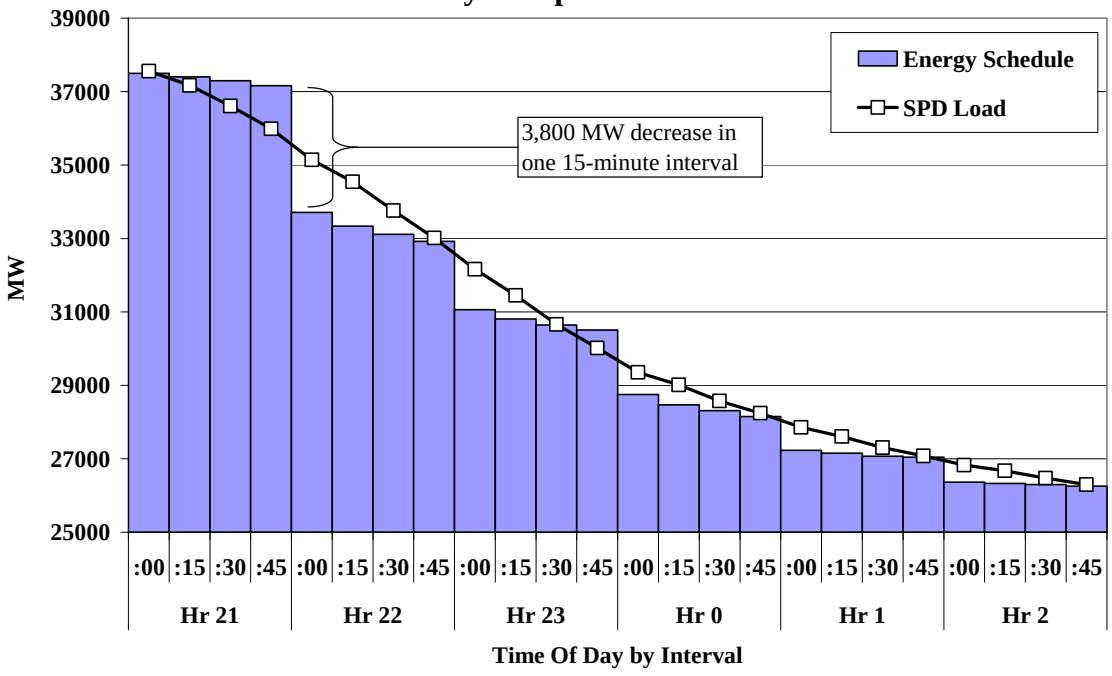
Figure 28: Final Schedules during Ramping-Up Hours
January to September 2004



On average, load increases from approximately 26 GW to more than 35 GW in the seven hours shown in Figure 28. The average increase per 15-minute interval is approximately 330 MW, although the rate of increase is greatest from 5:45 am to 7:00 am and relatively flat from 7:00 am to 8:30 am. The progression of load during ramping-up hours is steady relative to the progression of energy schedules. Energy schedules rise less smoothly, with small increases from the first to fourth interval in each hour and large increases from the fourth interval to the first interval of the next hour. For instance, the average energy schedule increases approximately 2,600 MW from the last interval of the hour ending 6 am to the interval beginning at 6 am, while the average energy schedule increases by several hundred megawatts in the subsequent three intervals.

The same scheduling patterns exist in the ramping-down hours. Figure 29 shows average energy schedules and load for each interval from 9:00 pm to 3:00 am during 2004.

Figure 29: Final Schedules during Ramping-Down Hours
January to September 2004



On average, load drops from more than 37 GW to approximately 26 GW in the six hours shown in Figure 29. The average decrease per 15-minute interval is approximately 480 MW, although the rate of decrease is greatest from 9:45 pm to midnight. The progression

of load during ramping-down hours is steady relative to the progression of energy schedules. As during the ramping-up hours, energy schedules decrease in relatively large steps at the top of each hour. For instance, the average energy schedule drops 3,800 MW from the last interval of the hour ending at 10:00 pm to the next interval.

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 QSEs submitting energy schedules that change hourly account for approximately one-half of the load in ERCOT, while QSEs that submit energy schedules that change every 15 minutes account for the other half of the load. Deviations between the energy schedules and actual loads (i.e., net balancing up energy equals real-time load minus scheduled energy) will result in purchases or sales in the balancing energy market.

Hence, Figure 28 indicates that during ramping-up hours, QSEs tend to purchase balancing energy on net at the end of each hour and sell balancing energy at the beginning of each hour. On the other hand, Figure 29 indicates that during ramping-down hours, QSEs tend to sell balancing energy on net at the beginning of each hour and purchase balancing energy at the end of each hour.

To evaluate the effects of systematic over- and under-scheduling more closely, we analyzed balancing energy prices and deployments in each interval during the ramping-up period and ramping-down period. Figure 30 shows balancing energy prices relative to balancing energy deployments for the ramping-up hours. This figure indicates two key characteristics of the balancing energy market outcomes.

First, balancing energy prices are highly correlated with balancing energy deployments. Second, there is a distinct pattern of increasing deployments 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 result is inefficient balancing energy prices. These prices are relatively volatile and could result in erratic dispatch signals to the generators. Figure 31 shows the same analysis for the ramping-down hours.

Figure 30: Balancing Energy Prices and Volumes Ramping-Up Hours
January to September 2004

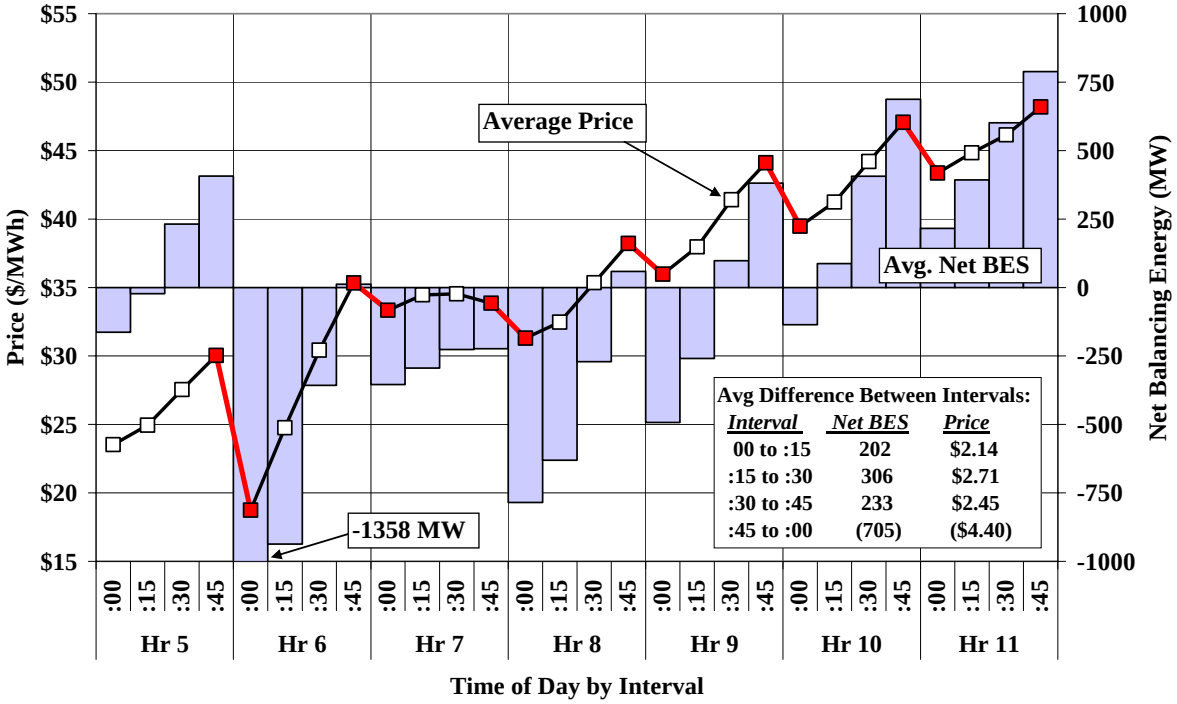
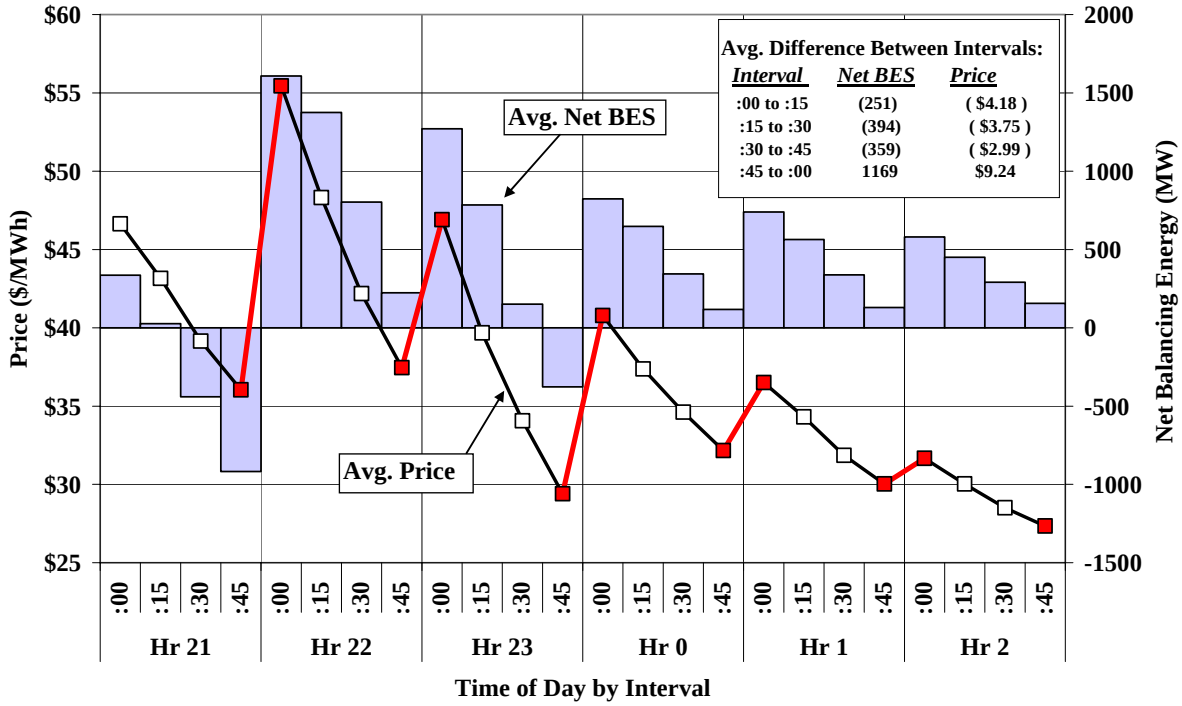


Figure 31: Balancing Energy Prices and Volumes Ramping-Down Hours
January to September 2004



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

While QSEs have the option to submit flexible schedules (i.e., every 15 minutes), most of the QSEs schedule only on an hourly basis, making little or no changes on a 15-minute basis. However, the two of the largest suppliers in ERCOT tend to schedule much more flexibly than other QSEs. These two large QSEs are virtually the only QSEs that submit schedules that vary by interval.

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 QSEs submit balancing energy bids and offers hourly and are made relative to the energy schedule. For example, if a QSE schedules 200 MW from a 300 MW resource, it may offer the remaining 100 MW in the balancing energy market. If it schedules 230 MW, it may offer 70 MW. However, if the energy schedule changes on a 15 minute basis, it may be difficult to reconcile the schedule with the hourly balancing offer, leading most QSEs to simply submit hourly schedules. This places a burden on the balancing energy market to reconcile the differences between the hourly schedules and the 15-minute actual load levels, which can result in inefficient price fluctuations.

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

Second, we recommend that ERCOT implement an optional capability for QSEs to automatically adjust their hourly balancing energy offers for the changes in their 15-minute schedules. This adjustment would assume that intra-hour increases in energy

schedules are supplied from the lowest-cost portion of the QSE's balancing energy offer. For example, assume a participant scheduled 600 MW in the first interval of the hour and offered 400 MW of balancing energy. If its schedule ramped up to 800 MW over the subsequent intervals

This would help ensure that the participant's portfolio energy offer is consistent with its energy schedules when the energy schedule is changing each interval.[18] These changes would likely increase the portion of the load that is scheduled flexibly and improve the performance of the balancing energy market.

B. Real-Time Operations and System Control

1. Regulation Need in ERCOT

The balancing energy market model runs an auction for power every 15 minutes. However, the system must remain in balance continuously. ERCOT uses regulation resources to adjust output every four seconds in order to keep load and supply balanced. In an electricity grid, any difference between load and supply causes the frequency on the system to deviate from the ideal level of 60 Hz. The frequency rises above 60 Hz when supply exceeds load, and frequency declines below 60 Hz when load exceeds supply.

Because ERCOT is separated from the eastern and western interconnects by DC ties, any over-generation or under-generation will translate directly into noticeable frequency deviations. This is not the case for control areas that are within the eastern interconnect where the inertia of hundreds of thousands of MWs of spinning generation will provide enough response to overcome a trip of one or two large units. Hence, the deployment of regulation plays a particularly critical role in ERCOT for controlling the system and maintaining reliability.

The regulation need is the amount of regulation that would be necessary to keep the system perfectly balanced at 60 Hz. When regulation need is positive, it implies that regulating units must increase output to keep frequency at the ideal level and vice versa.

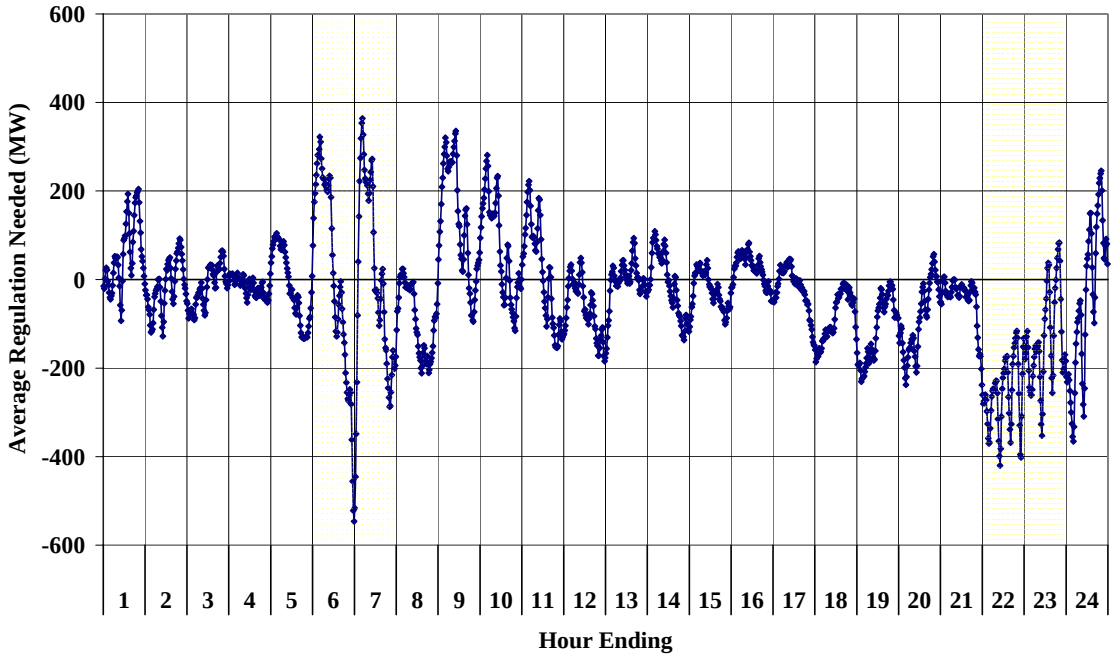
[18] Given how resource plans are used by ERCOT, we do not believe it would be necessary to allow resource plans that change each 15-minute interval.

ERCOT carries as much as 1,200 MW of regulation-up and 1.200 MW of regulation-down capability.

The actual regulation deployment usually does not precisely equal the regulation need, generally because: (a) the regulating units do not always accurately respond to the regulation signals; (b) ERCOT has exhausted its regulation capability, or (c) ramp rate constraints limit how much of the capability can be deployed. The difference between the regulation deployment and the regulation need is the area control error. Positive ACE occurs when generators are over-producing and negative ACE indicates that generators are under-producing. This sub-section of report evaluates patterns in regulation need and the next subsection examines the ACE by time of day.

Figure 32 shows the average regulation need for each minute of the day from January to September 2004. Regulation capability includes both regulation up (the ability to increase output when load exceeds generation) and regulation down (the ability to reduce output when generation exceeds load). In figure 5, the positive regulation needs indicate periods when regulation up is deployed, while negative regulation needs indicate periods when regulation down should be deployed.

Figure 32: Regulation Need by Time of Day
January to September 2004



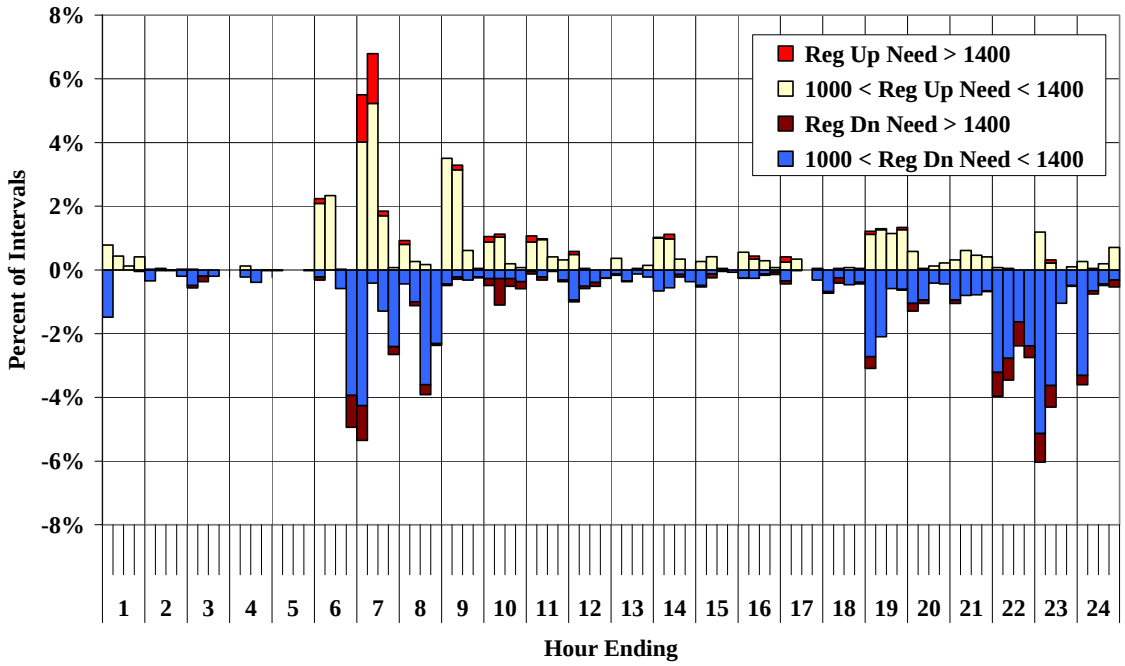
Regulation need fluctuates significantly throughout the day from a maximum regulation down need of 546 MW at 5:59 am to a maximum regulation up need of 364 MW at 6:11 am. Since these values are averages calculated over a 9 month period, the regulation need at a particular point in time may be substantially higher or lower than the average value shown. However, the figure shows that the regulation need tends to be predictably larger and more volatile during the morning load pick-up and evening load drop-off. All of the regulation up needs that average more than 300 MW occur between 5:00 am and 9:00 am, whereas the regulation down needs averaging more than 300 MW near 6:00 am and at various times between 9:00 pm and midnight.

The average regulation need during the morning pick-up has an hourly cyclical pattern such that the regulation need rises in the first half of the hour, falls during the hour, and reaches the hourly minimum in the last 15 minutes of the hour. The average regulation need during the evening drop-off has a cyclical pattern that rises and falls every fifteen minutes. Starting near 9:00 pm, the average need for regulation down increases significantly, suggesting that QSEs tend to over-produce during these hours.

While the average regulation need highlights systematic variations, the variation around the average is large. The frequent occurrences where ERCOT has to deploy large amounts of regulation up or regulation down has a greater impact on how much regulation ERCOT must procure than the average need.

Figure 33 shows a summary of how frequently the regulation needs were relatively large by interval. The bars shown in Figure 33 indicate the percent of one minute periods in each interval when the regulation need exceeds specific levels. As expected, ERCOT needed to deploy the largest amounts of regulation up and regulation down during the load pick-up and drop-off periods.

Figure 33: Percent of Intervals with Large Need for Regulation
January to September 2004



The figure shows that extreme quantities of regulation up are most frequently needed between 6:00 am and 6:30 am. Almost two-thirds of the instances when ERCOT's regulation up needs exceed 1,400 MW occurred in this 30-minute period. During the intervals ending 6:15 am and 6:30 am, ERCOT's regulation up needs exceed 1,000 MW 5.5 percent and 6.7 percent of the time, respectively. ERCOT needed more than 1,400 MW of regulation up approximately 1.5 percent of the time in both intervals.

Figure 33 also indicates that ERCOT's largest regulation down needs occurred during the morning pick-up and evening load drop-off. For the intervals ending 6:00 am and 6:15 am, the regulation down need exceeded 1,000 MW between 5.0 percent and 5.5 percent of the time while it exceeded 1,400 MW in both intervals approximately 1 percent of the time. During the evening load drop-off, ERCOT regulation down needs peaked in the interval ending 10:15 pm at 6 percent of the intervals.

The quantity of regulation up and regulation down procured in the day-ahead auction varies over the day, generally matching the pattern of regulation needs in Figure 33. Close to 1,000 MW of regulation up and 1,000 MW of regulation down is usually purchased during the morning load pick-up and evening load drop-off. The quantity of

regulation purchased is based on the most extreme needs rather than the average regulation needs. ERCOT staff has asked ERCOT stakeholders to consider an option where QSEs would provide larger amounts of regulation during the morning pick-up and drop-off in order to maintain frequency at these times.

The most frequent needs for very large quantities of regulation occur close to 6 am and 10 pm when the market participants in ERCOT are making the largest changes in their scheduled energy that forces the balancing energy market to function at its limits. The current market design envisaged balancing energy deployments as a residual for natural fluctuations in load and generator performance. QSEs, however, are providing energy schedules that are often inconsistent with actual increases and decreases in load in the early morning and late evening hours, forcing ERCOT to deploy large and fluctuating amounts of balancing energy. These issues, as well as the extent to which large needs for regulation deployments are leading to problems controlling the system (as evidenced by the ACE) are analyzed and discussed later in this section.

2. Evaluation of the Operator's Offset

One important tool the operators use to minimize the regulation deployments and ACE is the offset that will affect the balancing energy market deployments by adjusting the SPD load. Hence, when SCE is large and causing large regulation deployments, operators may adjust the offset to cause the balancing energy deployments to account for the SCE. This section evaluates the quality of the offset values being used by ERCOT.

Approximately 15 minutes prior to each interval and immediately before the SPD runs, the operator determines the offset which is added to the short-term load forecast to get the SPD load for the interval. The operator uses the offset primarily to minimize the need for regulation during the upcoming 15-minute interval. Currently, the offset is determined in a manual process every 15 minutes where operators try to forecast the actual load, market-wide SCE, and other factors that contribute to regulation needs.

To evaluate the process ERCOT operators use to calculate the offset, we compare the actual offset values to offset values derived from a simple offset methodology. Since the offset is intended to minimize the need to deploy regulation, a simple methodology for

determining the offset would be to adjust it by the inverse of the actual regulation deployment level in the current interval. Hence, if ERCOT is deploying 300 MW of regulation up in the current interval with an offset of zero, the offset for the next interval could be increased to 300 MW to cause the balancing energy market to increase energy deployments such that the need to deploy regulation up will be eliminated. ERCOT's method of calculating the offset is more sophisticated than this because other factors that can be forecasted will affect the need to deploy regulation in the next interval.

Hence, we can evaluate ERCOT's offset methodology by determining whether it is superior to the simple offset methodology described above. We do this by comparing the regulation needs with the actual offset versus what the regulation needs would have been using the simple methodology. The results of this analysis are shown in Figure 34 below.

Figure 34: Regulation Need from Actual Offset vs. Alternative Offset
1-Minute Averages – January to September 2004

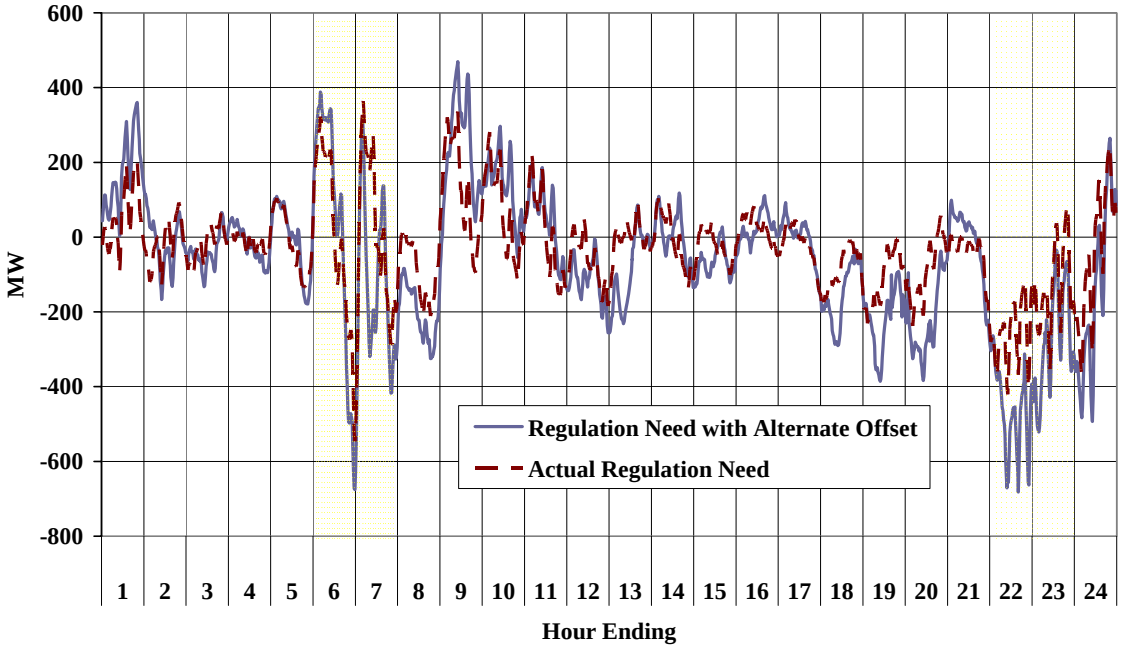


Figure 34 shows that for the majority of the day, the average actual regulation need resulting from the ERCOT's offset methodology is closer to zero than the average regulation need that would have occurred using the simplified benchmark methodology. This is particularly true in the afternoon and evening hours. However, in the period from 6 am to 7 am, the average regulation need from the simplified methodology was of

similar magnitude in the downward direction as the average actual regulation need was in the upward direction.

Due to the manual nature of the current offset calculation, it is difficult to assess the overall accuracy of the calculation. There are many uncertainties that must be taken into account when the operator determines the offset. If the operator's prediction of real-time factors affecting regulation needs is not perfect, the offset can actually exacerbate the need for regulation in some cases. However, this analysis demonstrates that, on balance, ERCOT operators are able to substantially reduce the need for regulation by considering other factors in calculating the offset beyond simply the current level of regulation deployments. ERCOT is working on a procedure to automate the calculation of the offset and the SPD load in order to improve transparency of the process and to rely less on the judgment of individual operators.

3. Area Control Error

Regulation providers receive dispatch signals from ERCOT every four seconds in order to maintain system frequency at 60 Hz. While regulation providers greatly improve the frequency, the frequency ordinarily deviates from the ideal level. When generation levels are higher or lower than the total system needs (load plus system losses), the frequency will deviate from the ideal level. Generally, every 0.01 Hz of frequency deviation corresponds to approximately 50 MW of ACE (either over-generation or under-generation). For example, if the total output in ERCOT is 39,900 MW (including regulation) when the load and losses total 40,000 MW, the ACE is equal to -100 MW and the frequency will equal 59.98 Hz.

The frequency can drop to 59.93 Hz before the operators are required to take actions under most circumstances, which corresponds to an ACE of 350 MW. Responsive reserves will be deployed when frequency drops to 59.91 Hz (i.e., 450 MW of under-generation). LaaRs that provide responsive reserves through UFRs are required to set the relays to trip when the frequency drops to 59.7 Hz at a minimum (i.e., 1500 MW of under-generation). Most of the UFRs are set at this level with the exception of one UFR with a load of 160 MW that is set to trip at 59.8 Hz. Because these settings require a

substantial amount of under-generation before the UFRs will trip and because ERCOT will take other actions as the frequency begins to fall, UFRs are rarely tripped.

Our first analysis of system control issues is in Figure 35, which shows the average ACE by time of day for the period from January through September 2004. These averages are calculated for each minute of the day.

Figure 35: Average ACE by Minute
January to September 2004

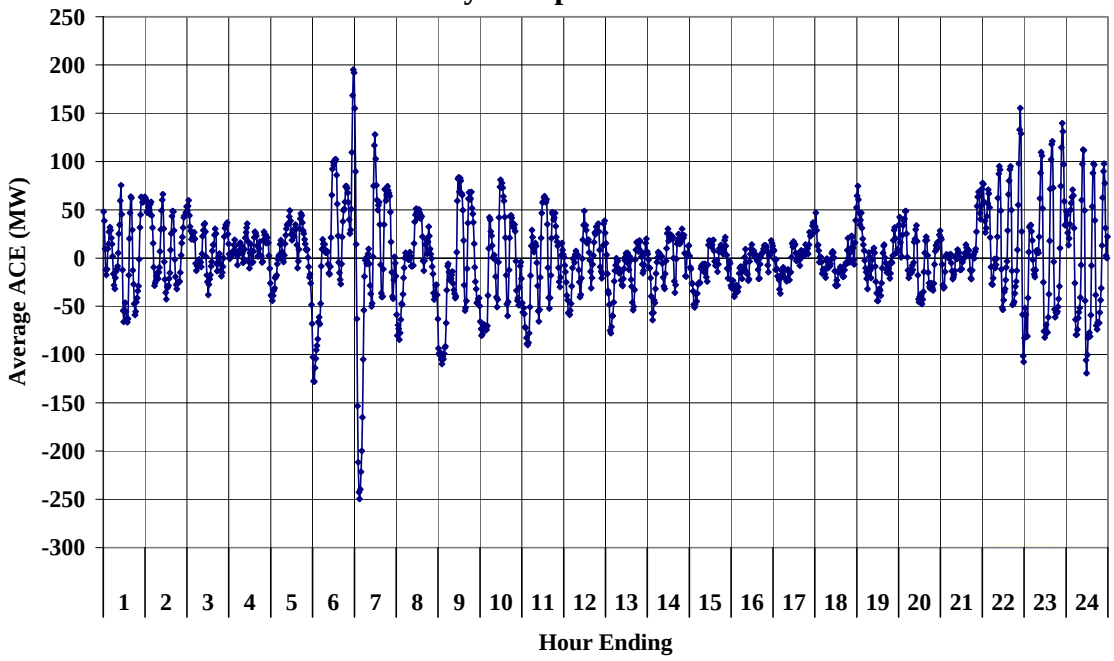


Figure 35 indicates that ACE fluctuates significantly throughout the day. Like the regulation needs analysis in the prior sub-section, the largest fluctuations occur in the morning load pick-up period and in the evening load drop-off. In fact, the largest positive and negative fluctuations occur shortly before and after 6:00 am. The maximum average ACE of 195 MW (i.e., over-generation of 195 MW) occurred at 5:58 am and the minimum average ACE of -250 MW at 6:07 am.

Since these values are averages calculated over a 9 month period, the ACE in a particular instance may be much higher or lower than the average shown in Figure 35. Hence, instances of relatively high ACE have occurred in many different times of day due to unexpected events (e.g., a generation or transmission line outage), however the ACE is

predictably volatile during the morning load pick-up and evening load drop-off. All of the averages greater than 100 MW or less than -100 MW occur during these times.

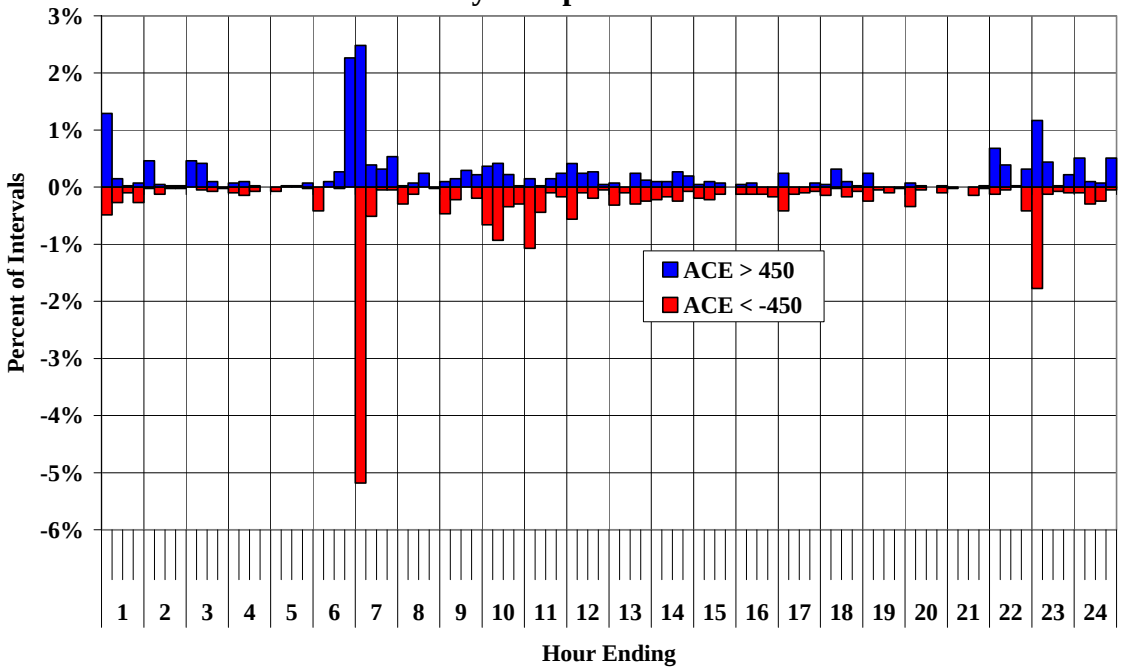
It is not surprising that the ACE fluctuations occur in the same periods as the fluctuations in regulation needs. When generation substantially exceeds load, ERCOT must deploy large amounts of regulation down, as it frequently does at 5:59 am. When QSEs do not fully respond to the regulation down deployments, there will be residual over-generation that is reflected in positive ACE. Likewise, when ERCOT has to deploy large amounts of regulation up shortly after 6:00 am, the ACE tends to be large and negative.

Another pattern that can be observed in Figure 35 is that the average ACE goes through four fluctuations each hour. While the amplitude of each fluctuation varies, these fluctuations roughly correspond to the length of 15-minute intervals. Every 15 minutes, the operator enters load into SPD that is intended to update portfolio dispatch levels in order to match generation and load, reducing the net regulation and ACE to zero. We will examine this 15-minute dispatch cycle in more detail later in this section.

Small frequency deviations have little impact on the system and are not a significant concern to ERCOT operators. However, large frequency deviations can have significant impacts, e.g., having to deploy responsive reserves that would otherwise be available to respond to system contingencies, such as generation outages. Hence, identifying how frequently the largest instances of ACE occur and when they tend to occur is important.

Figure 36 shows how frequently the ACE is greater than 450 MW or less than -450 MW which approximately correspond to frequency deviations of 0.09 Hz. ERCOT begins to deploy responsive reserves when there is a downward frequency deviation of 0.09 Hz or more.

Figure 36: Percent of Intervals with ACE Higher than 450 MW
January to September 2004



The bars shown in Figure 36 indicate the percent of one minute periods in each interval when the ACE is above 450 MW. The ACE rises above the 450 MW level (i.e., significant over-generation) most frequently during three periods: just after midnight, from 5:45 am to 6:15 am, and just after 10 pm. There were a large number of negative ACE events (i.e., significant under-generation) between 6:00 am and 6:15 am. More than 5 percent of the time within this interval, the ACE exceeded -450 MW (in absolute terms). A significant number of low frequency events occurred during the interval just after 10 pm as well.

When we calculated the same percentages on a daily basis rather than on an interval basis, we found that 41 percent of the days during the study period exhibited at least one instance of ACE lower than -450 MW and 38 percent of the days had one instance of ACE greater than 450 MW. These results, together with the results in Figure 36, are troubling because it suggests that ERCOT frequently has difficulty controlling the frequency of the system for short periods of time. The next sub-section evaluates factors that contribute to the need for regulation and the significant fluctuations in the system ACE.

4. Schedule Control Error and Load Deviations

Under the market protocols and operating procedures, ERCOT is responsible for accurately estimating load and losses and deploying generation sufficient to meet the total needs of the system. In this report, we refer to the difference between SPD load and actual load as the load deviation. This difference between instructed generation and actual generation is schedule control error, SCE. Positive SCE implies over-generation relative to the instruction. SCE includes any differences between actual output and instructions for regulating units.

ACE is primarily a function of the SCE and the load deviation. Positive SCE contributes to positive ACE, but it can be offset by negative load deviations. For example, if generators over-produce by 100 MW (i.e. SCE is 100 MW) and SPD load is below the actual load by 80 MW (i.e. load deviation is -80 MW), then the total ACE will equal 20 MW. This example shows that load deviations can be caused by offset values entered by the operator to counter the SCE. Therefore, load deviations do not necessarily imply an inaccurate short-term load forecast. To examine the patterns of SCE and load deviations, Figure 37 shows the average SCE and aggregate load deviations by time of day.

Figure 37: Schedule Control Error and Load Deviations
1-Minute Averages – January to September 2004

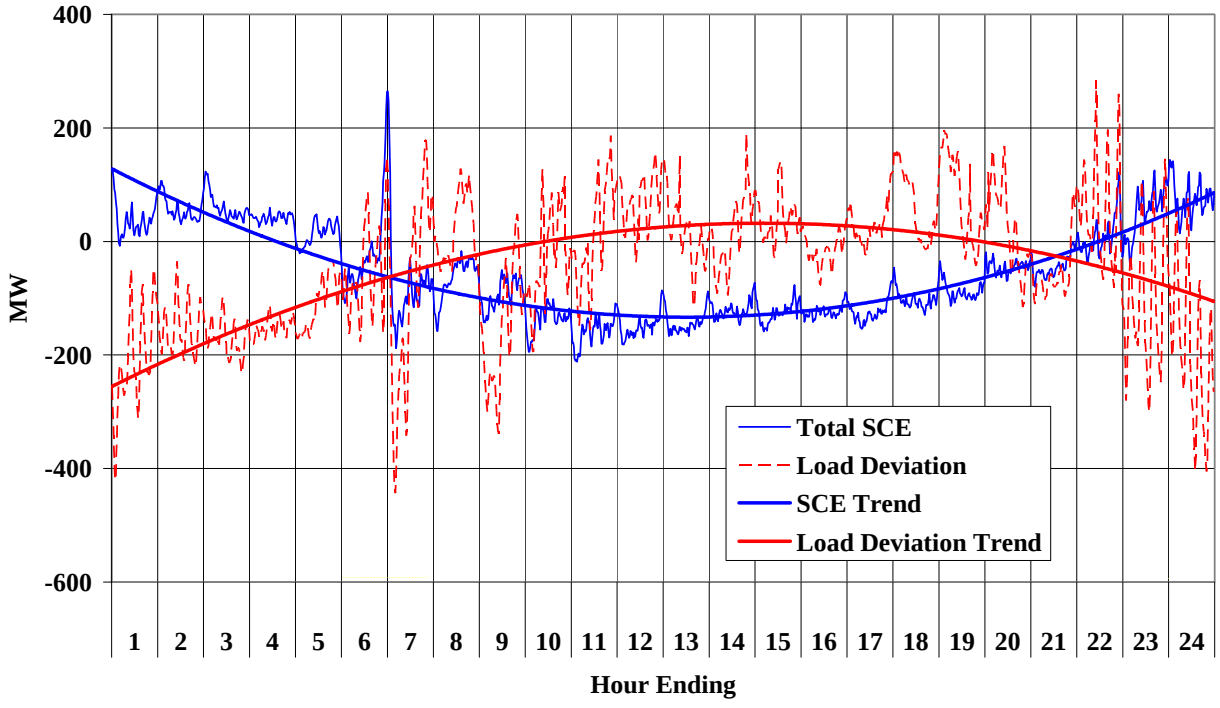


Figure 37 shows that the sum of SCE from all QSEs systematically moves in a distinct pattern. Before 5:00 am, generators tend to over-produce. However, generation deployments increase more quickly than actual output at 5:00 am, resulting in underproduction and a negative SCE. Immediately before 6:00 am, generators accelerate their output much more quickly than deployment instructions, causing the average SCE to be relatively large and positive. Shortly after 6:00 am, the average SCE becomes substantially negative.

Although the SCE tends to fall at the start of each hour and slowly rise during each hour in the morning and afternoon, the average market-wide SCE remains negative until the evening. Throughout the evening, the SCE trends upward until the average crosses into positive territory between 9:00 pm and 10:00 pm. During the evening load drop-off, there are generally no fluctuations comparable to the fluctuations during the morning load pick-up. The overall trend in the average SCE is shown by the quadratic trend line.

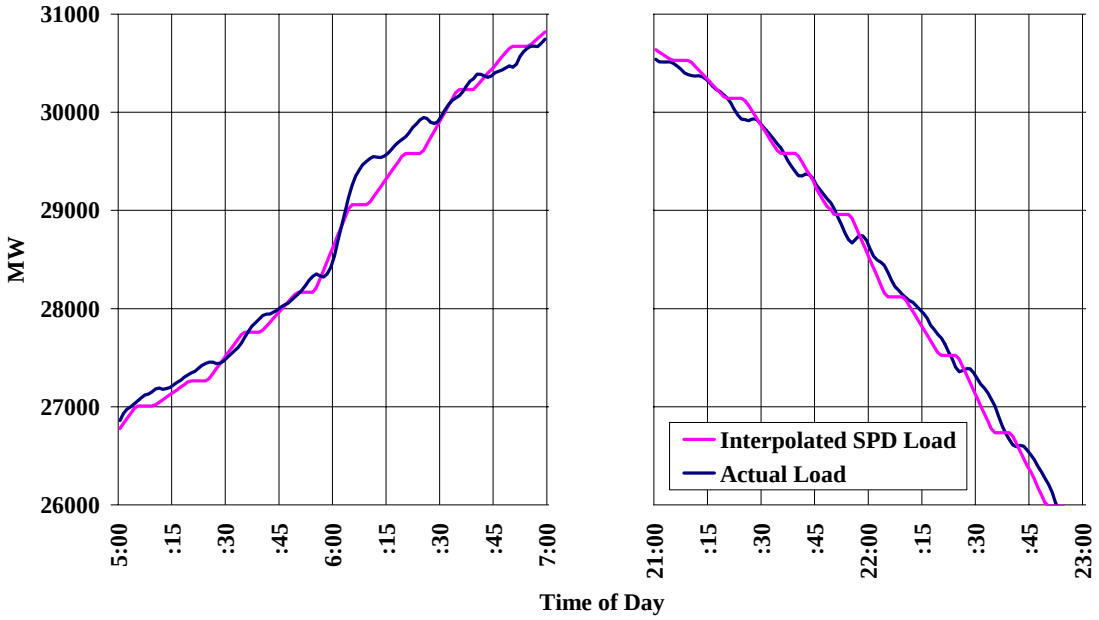
Figure 37 also shows a similarly distinct pattern in the load deviations. The load deviation is negative in early morning, trends upward until the afternoon when it turns and trends downward until the end of the day. In general, this pattern indicates that the load deviations tend to offset the trends in the SCE over the course of the day. This result can be explained, in part, by the fact that the load deviations include the effects of the offset. Since the offset is adjusted to reduce the need for regulation, it will be adjusted when SCE occurs that creates the need for regulation deployments.

We also find that the amplitude of fluctuations in the average load deviation is generally greater than the amplitude of the average SCE fluctuations. This may be due to the fact that the modeled load is assumed to be fixed for the middle 5 minutes of the 15-minute interval, while generators have the ability to adjust their output over the interval. The generation instructions (and the SCE) recognize that generators will take some time to move from the prior intervals output level to the current interval's output level.

In contrast, the operators submit a load quantity to SPD for each 15-minute interval that is used to calculate the balancing energy deployments. Because actual load does not move in a step-wise pattern from interval to interval, this process can lead to significant

load deviations. This cause of the load deviations can be understood by comparing the change in the actual load and SPD load on a one-minute basis during the ramp-up and ramp-down hours, which is shown in Figure 38. Seasonal versions of this figure are provided in Appendix A.

Figure 38: SPD Load vs. Actual Load during Ramping Hours
1-Minute Averages – January to September 2004



For the purposes of this figure, we show the single SPD load for each interval as the load for the middle 5 minutes of the 15 minute interval. In the last 5 minute period in the interval and the first 5 minute period of the following interval, we show the load moving in a linear manner between the SPD load levels in the two intervals. This pattern for the SPD load is consistent with the generation instruction provided by SPD, which assumes generators will be ramping during the first and third 5 minute periods in the intervals.

Figure 38 indicates the load deviations can become relatively large when SPD load is assumed to be constant and the actual load is changing rapidly. For example, when SPD load is flat from 6:05 am to 6:10 am, the actual load increases by 274 MW on average, leading to a sizable load deviation. During hours-ending 22 and 23, we observe a similar trend in the opposite direction. From 9:50 pm to 9:55 pm, actual load decreases by 348 MW while SPD load remains fixed. ERCOT utilizes regulation deployments to reconcile these temporary deviations.

These load deviations can be exacerbated by the fact that actual load does not increase and decrease at consistent rates. For instance, the actual load moves at a rate of 17 MW per minute from 5:55 am to 5:59 am on average, then increases at a rate of 133 MW from 5:59 am to 6:06 am. The implication of this pattern is that even if the SPD loads perfectly match actual load at the beginning and end of an interval, the deviation can be considerable in the middle of the interval.

One of the reasons our analysis focuses on the ramping periods is that the system control issues are most significant during these periods. The analyses of regulation deployments and ACE in the prior sub-sections show that the intervals near 6:00 am and 10:00 pm are the most volatile. This is due to the fact that load is changing rapidly during this period, as are the QSEs' schedules because many have bilateral contracts that serve 16 peak hours from 6:00 am to 10:00 pm.

To examine the causes of the ACE and other system control issues that occur in the ramping periods, the next analysis focuses on the load deviations and SCE levels in these intervals. The results of this analysis, presented in Figure 39, show the SCE and load deviations by minute in hours 5, 6, 21, and 22. Seasonal versions of this figure are provided in Appendix A.

Figure 39: SCE and Load Deviations during Ramping Intervals
1-Minute Averages – January and September 2004

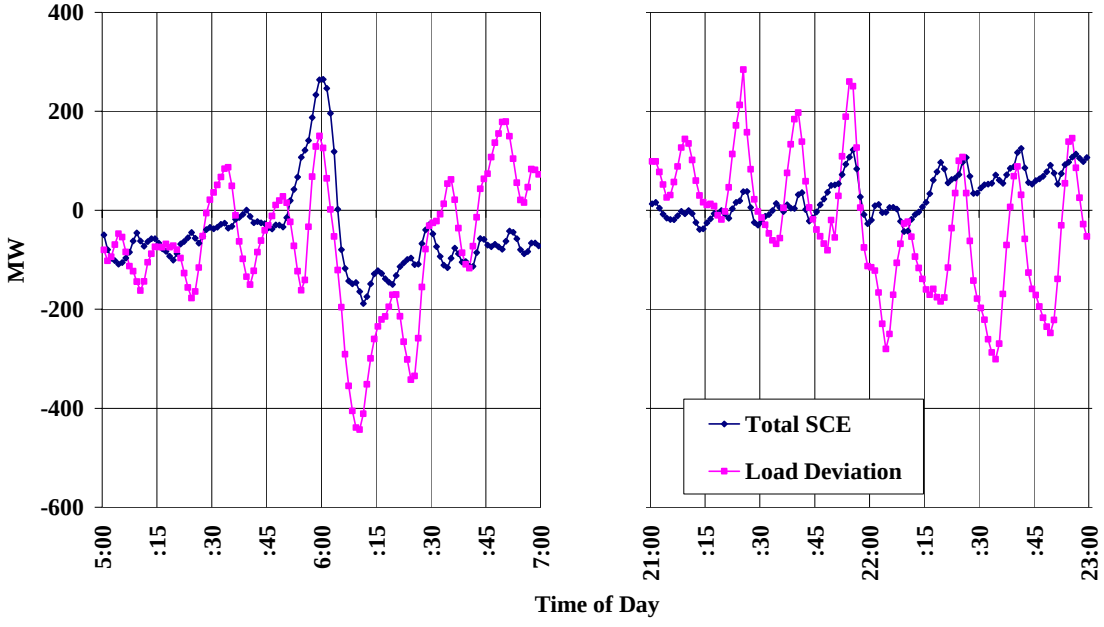


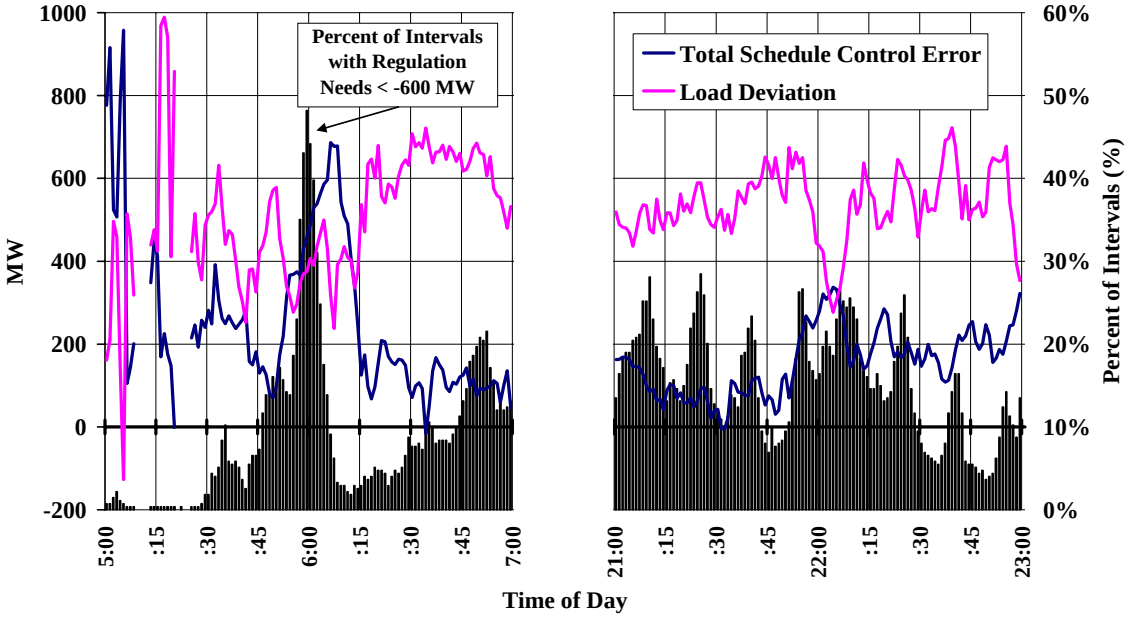
Figure 39 shows that in the middle 5 minutes of each 15-minute interval during the morning hours, the load deviation decreases sharply. This is consistent with the patterns shown in the prior figure, where the SPD load is fixed and actual load is increasing. Similarly, in the middle of each 15 minute interval during the evening hours, the load deviation increases sharply. The load deviations during the other portions of the intervals generally move in the opposite direction.

Because operators submit load to SPD every 15 minutes, they are limited in their ability to reduce the load deviations. The load deviation goes from -162 MW at 5:54 am to 150 MW at 5:59 am to -443 MW at 6:10 am. Making the average load deviation closer to zero at 6:10 am would necessarily require making the load deviation larger at 5:59 am.

Figure 39 shows that on average, SCE increases to 265 MW at 6:00 am and then falls to -188 MW at 6:11 am. This rapid adjustment in the average SCE consistently requires regulating units to increase output quickly. While Figure 37 above showed that SCE and load deviations tend to offset one another in general, they tend to move in the same direction around 6:00 am, jointly contributing to the relatively large need for regulation and ACE. This pattern does not hold close to 10:00 pm where the average SCE fluctuates much less. Fluctuation in the load deviations is the dominant contributor to the need for regulation and ACE that occurs close to 10:00 pm.

Our final analysis in this area shows the relative size of the load deviations and SCE in those intervals when the regulation need is relatively large. Figure 40 reports the average SCE and average load deviation during periods where the regulation need was lower than -600 MW in the four hours of the day that experience the most significant need for regulation.

Figure 40: SCE and Load Deviations in Periods with Large Negative Regulation Needs – 1 Minute Averages



The two lines in Figure 40 denote the average SCE and average load deviation in each minute when the regulation need was lower than -600 MW. The bars show how frequently the regulation need was lower than -600 MW for each minute during the four hours shown in the figure. The figure shows that the greatest need for regulation down was at 5:59 am when the regulation need was lower than -600 MW 49 percent of the time. This figure shows that the load deviations are generally larger than the SCE levels in these periods. Figure 41 shows the same analysis for periods when the regulation need was higher than 600 MW.

Figure 41: SCE and Load Deviations in Periods with Large Positive Regulation Needs – 1 Minute Averages

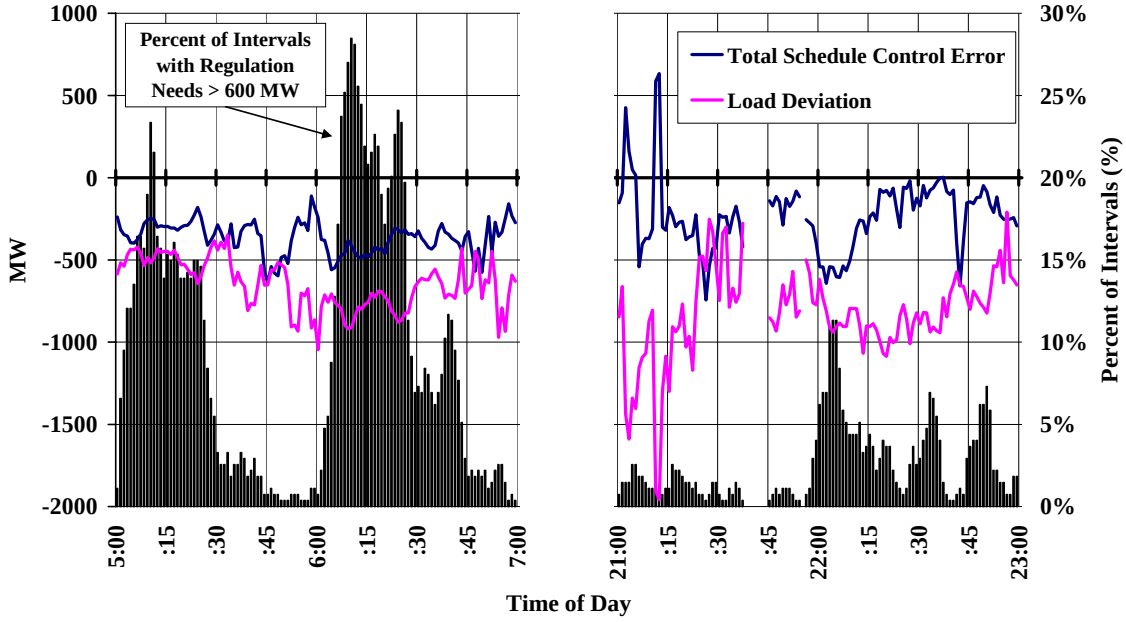


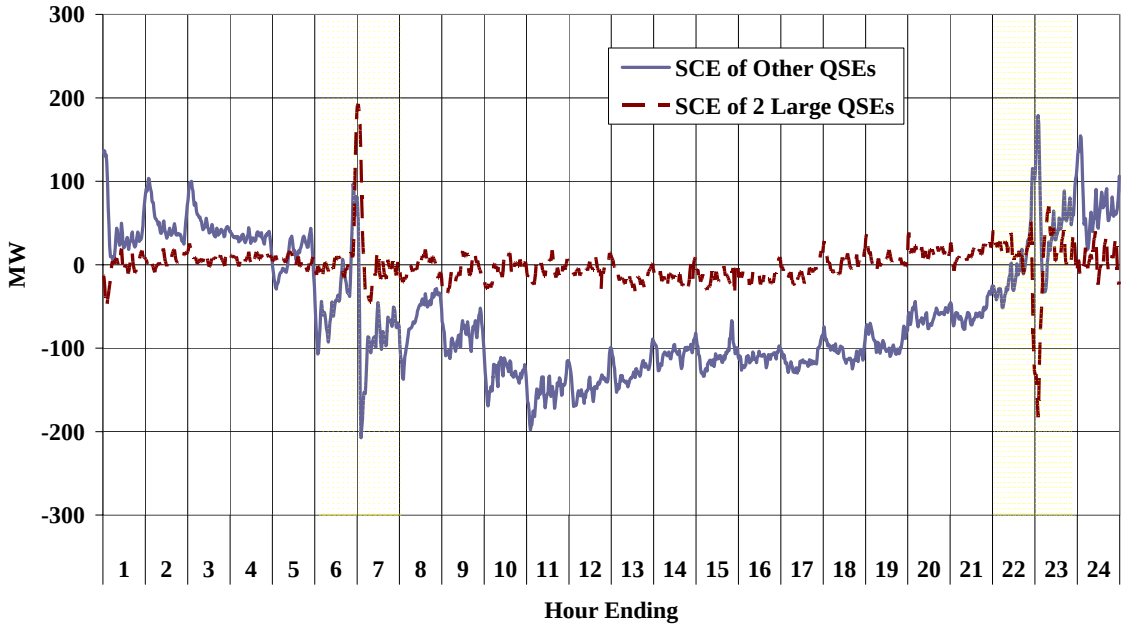
Figure 41 shows that the instances of large positive regulation needs are much more frequent in the morning ramp-up hours. The figure shows that the greatest need for regulation up was at 6:11 am when the regulation need was higher than 600 MW 28 percent of the time. On these days at 6:11 am, the average SCE was -460 MW and the average load deviation was -800 MW. This is consistent with the prior analysis, which showed that the load deviations tend to be a larger contributor to the need for regulation than the SCE, although the SCE is significant.

We make several recommendations at the end of this section to reduce the load deviations and SCE levels. First, we analyze QSE contributions to the SCE in the next subsection.

5. QSE contributions to the SCE

In order to better understand the variations in SCE that occur in ERCOT, the next several figures examine the SCE values for the larger and smaller QSEs. Figure 43 shows the average SCE in each minute of the day for the two largest QSEs (TXU and Texas Generating Company) compared with all other QSEs.

Figure 42: Total SCE for Large and Small QSEs
1-Minute Averages – January to September 2004



This figure shows that the two largest QSEs exhibit SCEs that are smaller and much less volatile on average, fluctuating in a narrow band around zero for most of the day. However, these QSEs exhibit an upward spike in their SCE around 6:00 am and a downward spike near 10:00 pm. Both spikes average 200 MW in absolute terms.

The other QSEs have positive SCE in the early morning and late night, and a persistently negative SCE for the rest of the day. Additionally, the SCE of the other QSEs tend to jump at the beginning of each hour. During the morning load pick-up and evening load drop-off, the SCE of other QSEs tend to peak at the top of the hour before decreasing sharply. Generally, this pattern results from the fact that all but the largest two QSEs change their schedules every hour on the hour.

The largest two QSEs account for approximately half of the energy that is generated in ERCOT. The previous figure shows that the largest two QSEs account for a relatively small portion of SCE that is systematically lower than zero in the afternoon. Figure 43 shows the average SCE for each QSE during afternoon hours in order to determine the relative contributions of each QSEs to the systematic under-generation.

Figure 43: Average SCE by QSE during Afternoon Hours
Noon to 6 pm – January to September 2004

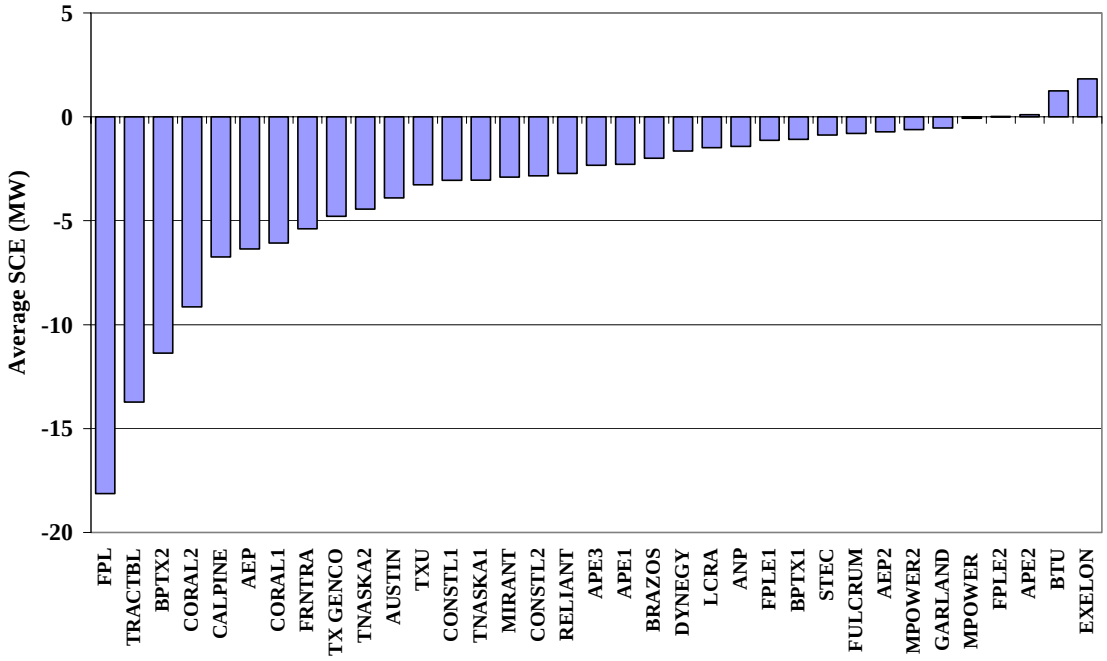


Figure 43 shows that a large majority of QSEs under-generate by a slight margin on average during the afternoon. Responsibility for systematic under-production is wide-spread, although the three worst QSEs averaged below -10 MW and accounted for 36 percent of the aggregate under-generation while the eight worst QSEs accounted for 63 percent of the aggregate under-generation. These QSEs have relatively large portfolios so that the systematic SCE is small as a percentage of their capacity. As a share of capacity Tractebel (2nd) and Frontera (8th) were the only two QSEs of the eight that had a systematic error of more than one percent -- 1.2 percent and 1.1 percent, respectively.

Because the SCE levels are most critical during the ramp-up and ramp-down hours, we show the same comparison of SCE levels for large and small QSEs in these periods in Figure 44.

Figure 44: SCE Levels for Large and Small QSEs in Ramping Hours
1-Minute Averages – January to September 2004

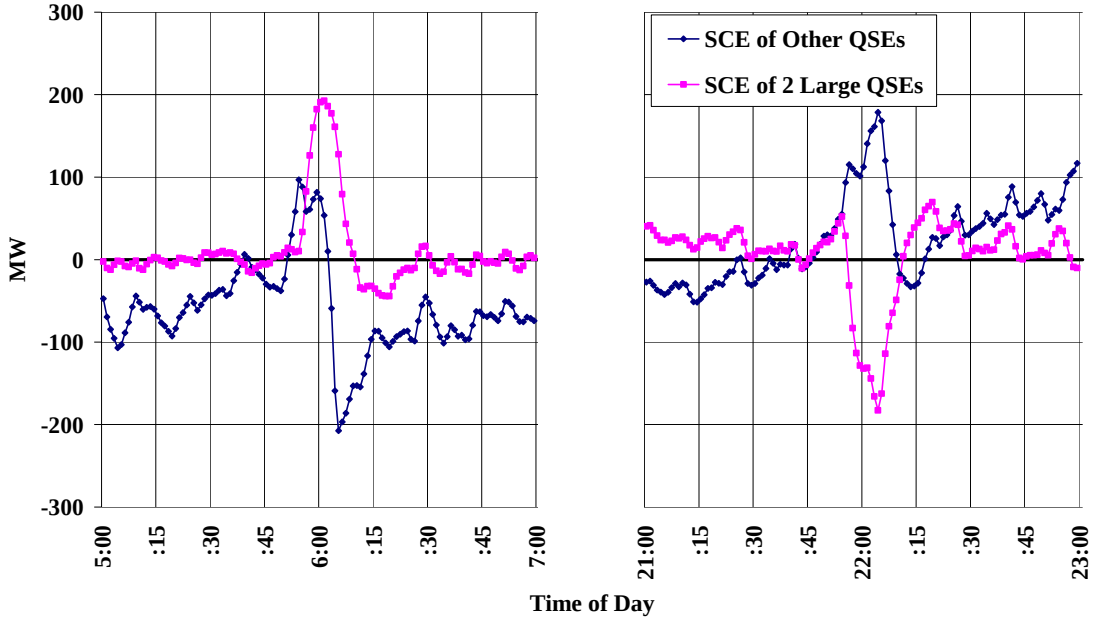


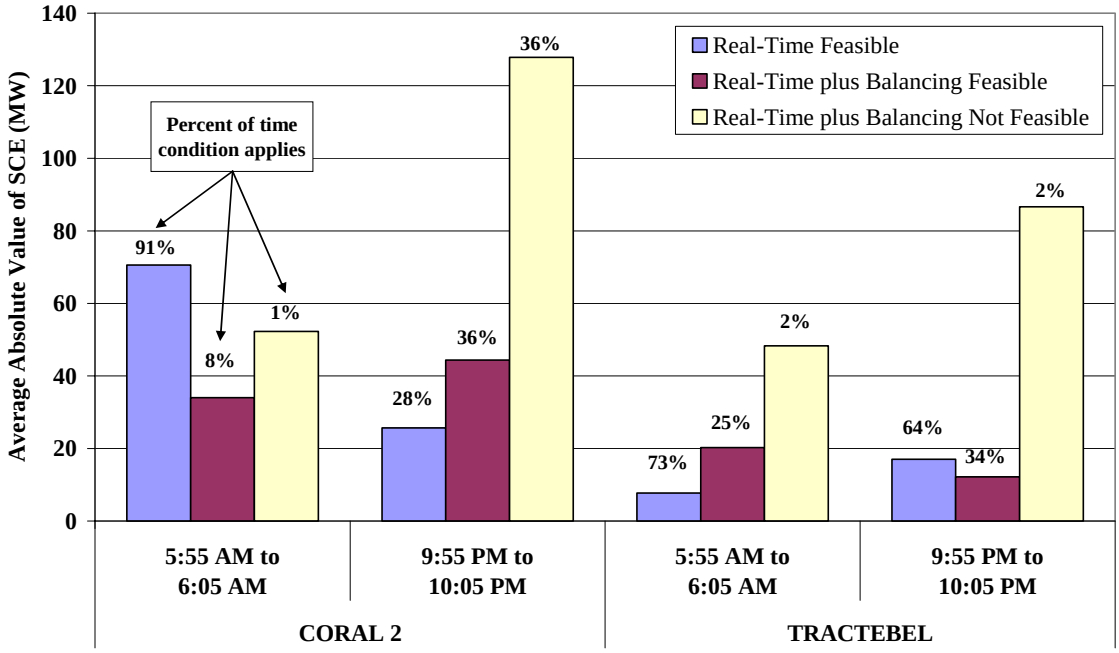
Figure 44 shows that the SCE of both groups of QSEs rises sharply in the interval preceding 6:00 am, before dropping more sharply in the interval after 6:00 am. Because these movements are in the same direction, they combine to increase the SCE fluctuation. These movements can be explained by the sharp increase in energy schedules that occur at 6:00 am. Many generators begin ramping prior to 6:00 am in anticipation of the large schedule change leading to a positive SCE. Once the sharp energy schedule changes occur around 6:00 am, the SCE decreases rapidly because the energy schedules increase much faster than the suppliers' output.

The intervals before and after 10:00 pm exhibit the opposite pattern. The fluctuations of the large and small QSEs in these intervals are in opposite directions and tend, therefore, to reduce the total SCE levels on average. This difference between the SCE patterns at 6:00 am and 10:00 pm are due to the differences in energy scheduling patterns of the two groups of QSEs. In particular, the large QSEs generally increase their energy schedules sharply in the interval from 10:00 pm to 10:15 pm. Because they do not alter their physical generation to precisely match this increase, it results in a negative SCE.

6. Infeasible energy schedules

Figure 28 and Figure 29 show that energy schedules increase abruptly at 6 am and step down sharply at 10 pm. In just one interval schedules rise an average of 2,600 MW at 6 am and decline 3,800 MW at 10 pm. The sharp scheduling changes coincide with significant SCE that contributes to frequency deviations. One concern is that QSEs may submit schedules that are physically infeasible and then be unable to obey dispatch instructions at 6 am and 10 pm. We reviewed energy schedules and resource-specific ramp rate information to determine whether QSEs submit physically infeasible energy schedules in real-time. Figure 45 summarizes the relationship between SCE and infeasible schedules for the two QSEs that submitted a significant number of infeasible schedules.

Figure 45: SCE vs. Feasibility of Real-Time and Balancing Energy Schedules Select QSEs – January to August 2004



The first of the three bars in Figure 45 show the average magnitude of SCE for each QSE when it submits a feasible real-time energy schedule. The next two bars are conditions under which the energy schedules are not feasible. When QSEs submit infeasible real-time schedules, they sometimes receive balancing energy deployments that cause the sum of the schedules and deployment to be feasible. These cases are shown in the second bar. The third bar shows the SCE when the sum of the real-time schedule and balancing

schedule are not feasible. The percentage on each bar shows how frequently each type of energy schedule occurred.

In general, the average SCE is at least 400 percent larger in cases when the total energy schedule is infeasible compared to when the real-time energy schedule is feasible. The exception is the case of Coral 2 in the morning transition, where infeasible scheduling was very infrequent. The SCE tends to be larger when the real-time schedules are infeasible, even when the balancing energy deployments cause the total changes in output to be feasible.

Overall, infeasible scheduling occurs on a limited basis and tends to be mitigated by balancing energy schedules. However, to the extent it occurs, infeasible scheduling tends to increase SCE during periods where the ERCOT system is particularly sensitive. We see no compelling reason to allow physically infeasible energy schedules. Some would argue that because the schedule changes are linked to bilateral contracts, they must occur at 6:00 am and 10:00 pm. This is not true. If the balancing energy prices are compensatory (i.e., higher than the production costs of the generator), then it is profitable to bring the unit online prior to 6:00 am and sell the excess power into the balancing energy market.[19] If prices are not compensatory, it would be more profitable for the generator to start-up its units after 6:00 am and purchase balancing energy to satisfy the bilateral obligation.[20]

We would note that the balancing energy prices from 6:00 am to 6:45 am are lower on average than the prices from 5:00 am to 6:00 am. Hence, starting units at 6:00 am cannot be the most profitable strategy for a supplier, regardless of their contract position.

---

[19] Units brought online prior to 6:00 am could be scheduled at their minimum generation level and offered into the balancing energy market at their marginal cost.

[20] This can be accomplished in one of two ways. First, if the generator also serves load, it could reduce its load schedule by an amount equal to the energy that would have been produced by the unit, resulting in a purchase from the balancing energy market as a load deviation. Second, the counter-party to the bilateral contract could reduce its load schedule until the unit is going to be brought online and a payment equal to the balancing energy price would be made by the supplier until that time.

7. Recommendations for Reducing Regulation Need and ACE

In this section, we identified two factors that contribute to the large fluctuations in the regulation needs and ACE analyzed earlier in this section: schedule control error and load deviations. Although both factors are significant, the load deviations are a larger contributor to the system control issues.

We find that the most significant cause of the load deviations is the load plateau that is modeled in the middle 5-minutes of each interval. Because generation is not expected to move during these periods, but actual load is moving, the regulating capability must move to match all changes in load during the periods. The deviations could be substantially reduced by:

- Eliminating the load and generation plateau in the middle of the interval by changing the modeling convention so that QSEs ramp to the generating level by the middle of the interval (7.5 minutes), then beginning to ramp immediately to the generation level for the next interval; or
- Eliminating the 5-minute plateau by reducing the interval length to either 5 minutes or 10 minutes. The shorter the interval, the less ERCOT will have to rely on regulation deployments within the interval.

The first of these two approaches would likely require fewer changes to the market models and settlement systems. This change would have the added benefit of making more capability available to the balancing energy market by lengthening the ramping periods.

With regard to SCE, we have two recommendations for ERCOT to consider that should reduce SCE levels:

- Require that QSEs submit physically feasible energy schedules. It would be difficult to automatically validate the schedules for feasibility since they can be submitted well before the operating hour. However, it could be monitored and enforced through ex post validation of compliance.
- Implement uninstructed deviation charges that allocate a portion of the regulation costs to the QSEs exhibiting large SCE. We are aware that PRR 356 has been pending since September 2002 that would implement this type of allocation. This PRR would be an improvement over the status quo, but the incentive effects would be improved by allocating the regulation costs as follows:

- Calculate the regulation needs (regulation deployments – ACE) and the SCE for each QSE on a one minute basis.
- Select the 15 minutes in each hour when the absolute value of the regulation needs is the highest.
- Charge QSEs the average of the absolute value of their SCE values in those 15 minutes.
- The residual regulation costs would be allocated on a load ratio basis.

The current uninstructed deviation penalties primarily address "price chasing" and a certain gaming strategy but do not provide meaningful disincentives for QSEs to minimize SCE that increase regulation needs and costs. The proposed allocation of regulation costs would focus on the periods in each hour with the highest regulation needs, which generally governs how much regulation is purchased, and allocate the costs in proportion to the magnitude of the SCEs in those periods. If SCEs are low, and thus not a large determinant of the need for regulation in those periods, the costs allocated to the generators in those periods will also be low.

B. Portfolio Ramp Constraints in SPD

In electric power markets, there are physical limits to the rate at which each resource can increase output. At times demand increases so rapidly that it is not possible for the least expensive unloaded resources to respond in economic merit order. Because resources are offered in the balancing energy market in portfolios, each QSE must submit a portfolio ramp constraint allows that QSE to manage how quickly the balancing energy market requests that it increase or decrease production from its portfolio.

Accurately representing ramp limitations is important so that the market produces efficient prices and dispatch instructions. This section examines the current method of reflecting ramp constraints in the balancing energy market model. Based on this examination, we identify some concerns and recommend an alternative approach.

Currently, ERCOT's dispatch model imposes ramp constraints on the balancing energy market solution to ensure that QSEs can comply with their dispatch instructions. Each QSE submits up-balancing offers for each of its zonal portfolios with an associated ramp

rate. Furthermore, each QSE also submits down-balancing offers with associated ramp rates. The model is limited by the following two ramp constraints:

1. NetUpBalanceMW(t) – NetUpBalanceMW(t-1) <= RampRate * 10 minutes

2. (-1) * RampRate * 10 minutes <= NetUpBalanceMW(t) – NetUpBalanceMW(t-1)

The ramp constraint limits the change in the net up-balancing energy cleared (NetUpBalanceMW) from interval t-1 to interval t. The first equation limits changes in the positive direction while the second equation limits changes in the negative direction. Figure 46 shows an example of how the ramp constraint might affect balancing energy deployments.

Figure 46: Effects of Portfolio Ramp Constraints

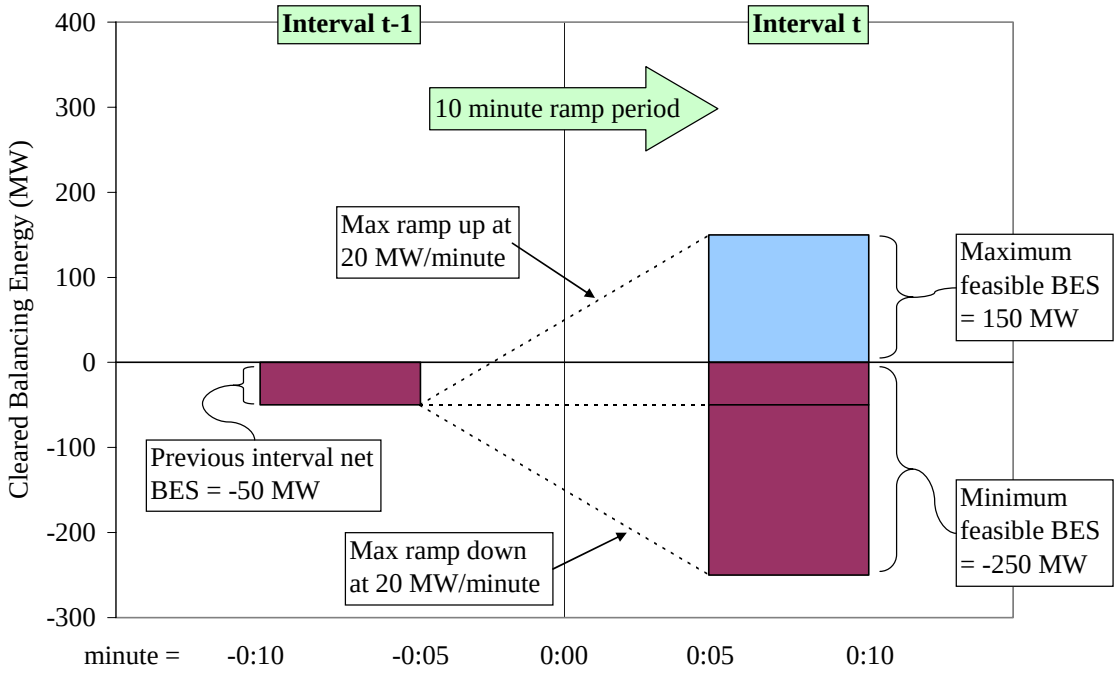


Figure 46 shows an example where a particular QSE's net balancing energy deployed is -50 MW in the previous interval and its portfolio ramp rate is 20 MW/minute. If the QSE increases at the maximum rate for 10 minutes, it is possible to increase output by 200 MW from the previous interval. Thus, the maximum feasible cleared quantity is 150 MW in the current interval. Likewise, if the QSE decreases at the maximum downward

rate for 10 minutes, it is possible to decrease output by 200 MW from the previous interval. Thus, the minimum feasible cleared quantity is -250 MW in the current interval.

The dispatch model does not take energy schedule changes into account when imposing ramp constraints. Presumably, QSEs know this when they submit portfolio ramp rates and alter their ramp rates to ensure that the resulting dispatch instructions are feasible. However, the market outcomes would be more efficient in many cases if energy schedule changes were taken into consideration when ramp constraints are imposed by the balancing energy model.

The interval ending 10:15 pm on September 9, 2004 provides a useful illustration of this. The following table shows selected information from interval-ending 10:15 pm and the preceding interval on September 9, 2004.

Table 6: Clearing Price, Load, Scheduled Energy and Balancing Energy Intervals-Ending 10:00 pm and 10:15 pm, September 9, 2004

			All QSEs		Single QSE	
Interval-Ending	**MCPE**	**SPD Load**	**Scheduled Energy**	**Net UBES**	**Scheduled Energy**	**Net UBES**
22:00	$30.21	38,600	39,669	-1,069	3,391	-110
22:15	$290.01	37,500	33,486	4,014	2,950	135
Change	$259.80	-1,100	-6,183	5,083	-441	245

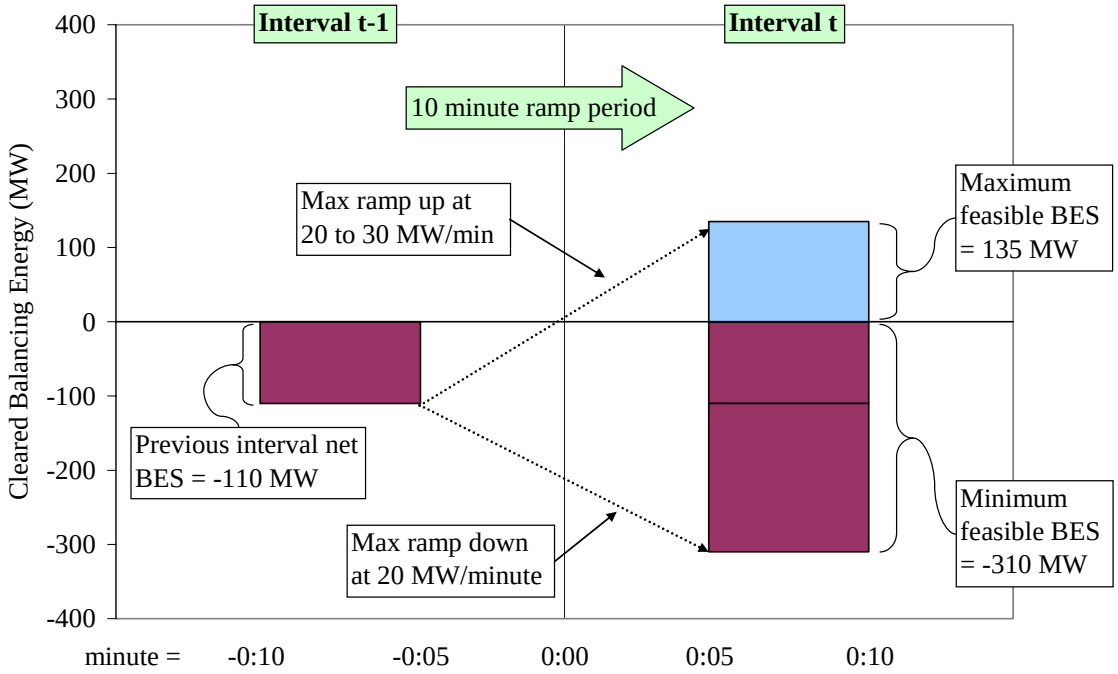
The dispatch model was not able to solve in interval ending 10:15 pm because of insufficient resources that could ramp up in one interval. Although, the price was revised by the Modified Competitive Solution Method to $109.17/MWh, an offer was accepted at $290.01 because less expensive resources could not be ramped quickly enough. The ramp constraint and resulting high price were caused by the net upward balancing MW change of 5,083 MW from interval ending 10:00 pm to interval ending 10:15 pm.

During this time, demand actually decreased by 1,100 MW so aggregate generation was decreasing. From interval ending 10:00 pm to interval ending 10:15 pm, the reduction in demand was far exceeded by the reduction in scheduled energy, which resulted in the

substantial change in cleared balancing energy. The rapid *increase* in cleared balancing energy results in binding ramp constraints. However, resources are physically *decreasing* their output levels. If the model took account of the decrease in schedules, the dispatch instructions would not have been constrained by ramping limitations and it would not be necessary to accept a balancing energy offered at $290.01/MWh.

Table 6 above includes selected quantities for a particular QSE to illustrate how the ramp constraint functioned in a particular case. From interval ending 10:00 pm to 10:15 pm, this QSE ramped up at the maximum rate allowed by the dispatch model. The QSE's ramp rate was 30 MW/minute in the up-balancing range and 20 MW/minute in the down-balancing range. So, the QSE was able to ramp from -110 MW to 0 MW in the first 5.5 minutes and from 0 MW to 135 MW in the subsequent 4.5 minutes. This is shown in Figure 47.

Figure 47: Portfolio Ramp Constraints
Interval-Ending 10:15 pm, September 9, 2004



From interval-ending 10:00 pm to 10:15 pm, the QSE shown in Figure 47 reduced its energy schedule from 3,391 MW to 2,950 MW. At the same time, this QSE's balancing energy deployments increased from -110 MW to 135 MW. Although this QSE had an

additional 400 MW of excess capacity offered at prices between $48/MWh and $85/MWh, ramp limitations prevented a change of more than 245 MW in cleared energy from the previous interval. Ironically, the output from this QSE changed very little during this period and its units were not physically ramp constrained.

As described above, the dispatch model currently imposes ramp constraints to prevent large changes in *balancing energy deployments* rather than large changes in *total output*. Figure 48 illustrates how QSE is likely to be affected by the current ramp constraint methodology when there are large energy schedule changes. The case shown in Figure 48 is the same QSE and interval as shown in Figure 47.

Figure 48: Portfolio Ramp Constraints incorporating Schedule Changes
Interval-ending 10:15 pm, September 9, 2004

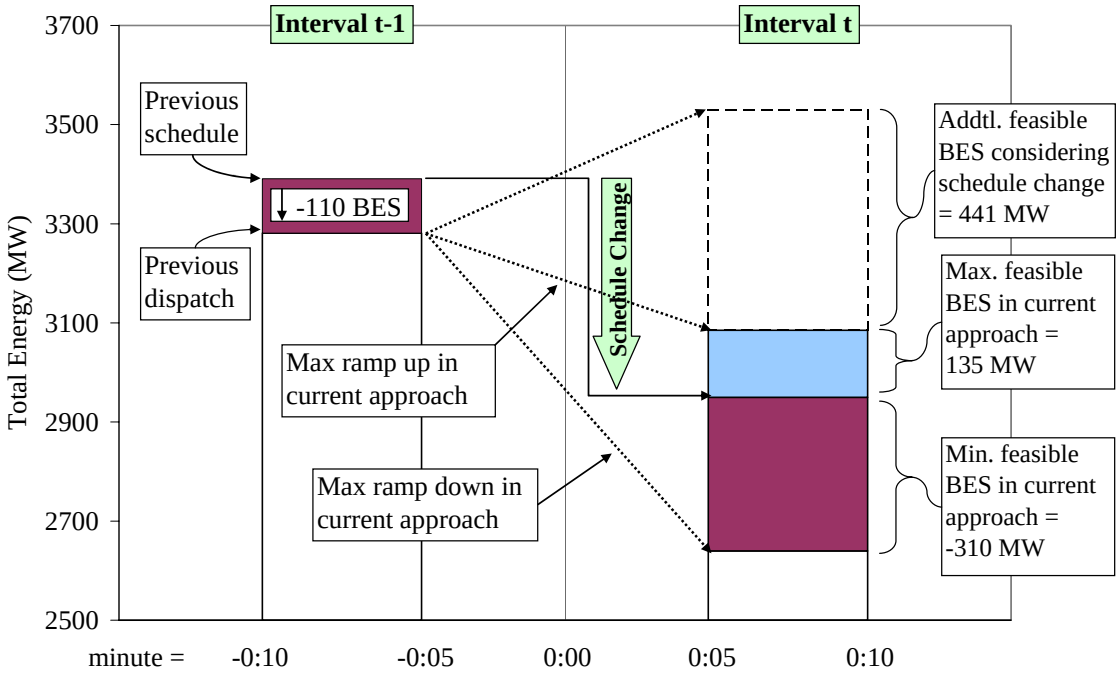


Figure 48 shows that in interval-ending 10:00 pm, the QSE scheduled 3,391 MW of energy and received balancing deployments of -110 MW for a total energy output of 3,281 MW. In the interval-ending 10:15 pm, it reduces its schedule by 441 MW to 2950 MW. Applying the current ramp constraint methodology that considers only the prior balancing energy deployment of -110 MW, the model may issue a minimum deployment

to the QSE of -310 MW (2640 MW total output) or a maximum deployment of 135 MW (3085 MW total output).

However, ramp rate limitations exist to reflect the fact that it takes time for physical units to change their output level. Therefore, such limitations would most naturally apply to the change in total output rather than the balancing energy deployments. In this example, the QSE reduction in scheduled energy, if recognized, would allow ERCOT to issue a balancing energy deployment of 576 MW rather than the 135 MW the current methodology would allow. This increase of 441 MW is enabled by the fact that ERCOT would simply be requesting that the QSE not follow the output reduction in its energy schedule. The extra balancing energy that is physically available, 441 MW in this case, would be make more supply available in intervals with large balancing energy demands and substantially reduce inefficient balancing energy price volatility.

Therefore, we believe a more accurate way to the SPD to model the portfolio ramp constraints would be to apply the ramp limitations to changes in *total energy output,* which would appropriately take account of changes to energy schedules and balancing energy deployments. This approach would prevent the balancing energy market from exhibiting artificial shortages when there are large changes in scheduled energy but no physical shortages. Under this formulation, the model would utilize the following two modified ramp rate constraints:

1. OutputMW(t) – OutputMW(t-1) <= RampRate * 10 minutes
2. (-1) * RampRate * 10 minutes <= OutputMW(t) – OutputMW(t-1)

   where: OutputMW(t) = SchedEnergyMW(t) + NetUpBalanceMW(t).

This ramp constraint would limit the net change in the *total output* from the QSE from interval t-1 to interval t. The first equation would limit changes in the positive direction while the second equation would limit changes in the negative direction. To illustrate the impact of reformulating the ramp constraint in the balancing energy market, we re-solved the dispatch model for the September 9, 10:15 pm interval with the alternate constraint.

Table 7: Clearing Price, Load, Scheduled Energy and Balancing Energy Using Re-formulated Ramp Constraint Intervals-Ending 10:00 pm & 10:15 pm, September 9 2004

			All QSEs		Single QSE	
Interval-Ending	MCPE	SPD Load	Scheduled Energy	Net UBES	Scheduled Energy	Net UBES
22:00	$30.21	38,600	39,669	-1,069	3,391	-110
22:15	$73.08	37,500	33,486	4,014	2,950	493
Change	$42.87	-1,100	-6,183	5,083	-441	603

Under the new ramp constraint formulation, the dispatch model was able to serve demand while observing ramp constraints and set a balancing energy price of $73.08/MWh. The demand, scheduled energy, and net up-balancing energy are the same as in Table 6, since none of these were affected by the altered ramp constraint. The QSE shown in Table 6 is also shown here with identical energy schedules. However, the net up-balancing energy is 493 MW; 358 MW higher than with the current ramp rate methodology.

Based on the analysis in this section of the report, we recommend that ERCOT consider modifying the portfolio ramp rate methodology as described above to appropriately take account of changes to energy schedules. This should prevent the balancing energy market from exhibiting artificial shortages when substantial quantities of rampable balancing energy are actually available.

C. QSE Provision of Reserves

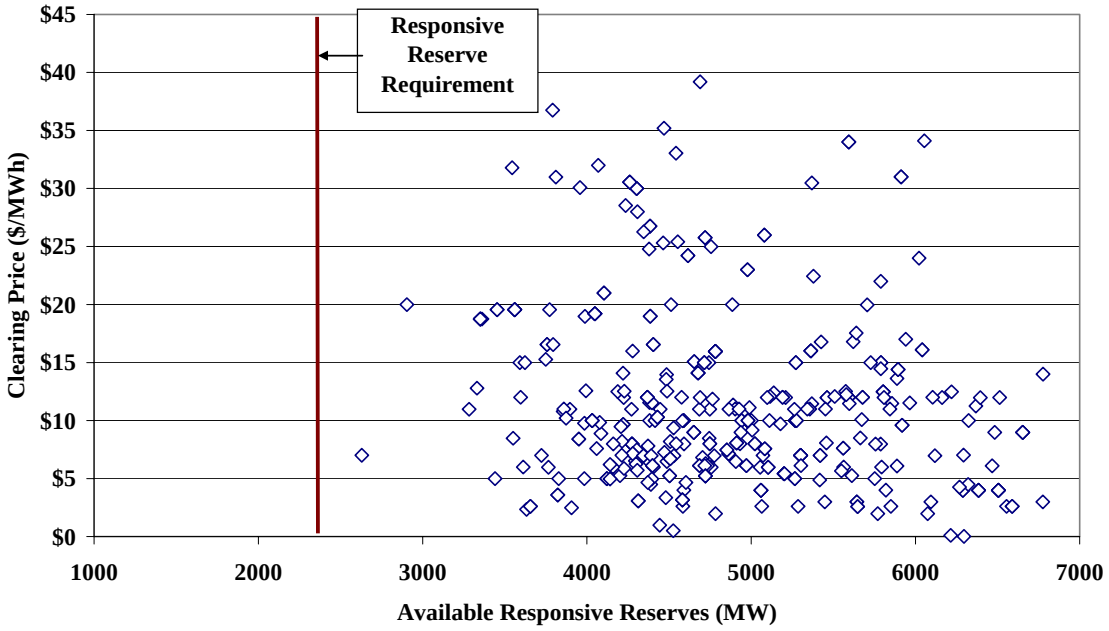
In order to maintain the reliability of the ERCOT system, ERCOT requires 2,300 MW of responsive reserves (i.e., 10-minute synchronous reserves). ERCOT satisfies this requirement through resources that QSEs self-schedule and by procuring responsive reserves through the ancillary services auction market run one day prior to the operating day.

QSEs responsive reserve schedules, either by self-scheduling resources or by clearing in the ERCOT responsive reserve market, are obligated to set aside sufficient capacity to respond to a reserve deployment. This sub-section of the report assesses the adequacy of

responsive reserves actually available in real-time during the first eight months of 2004 and evaluates whether QSEs are individually satisfying their obligations.

Figure 49 plots the hourly real-time responsive reserves capability against the responsive reserves prices in the peak hour of each day. The capability calculated for this analysis is based on the maximum capacity and actual energy output of each generating unit. Hence, units producing energy at their maximum capability will have no available responsive reserves capability. Although ERCOT requires loads to procure 2,300 MW of responsive reserves day-ahead, any partially loaded resource capable of receiving real-time dispatch signals can respond in the event of a contingency. Thus, there is usually a surplus of responsive reserves in real-time. The figure also shows the responsive reserves requirement of 2,300 MW to show the amount of the surplus in each hour.

Figure 49: Actual Responsive Reserves Capability and Clearing Price
Peak Load Hours – January to August 2004



This figure indicates a somewhat random pattern of responsive reserves prices in relation to the hourly available responsive reserves capability in real time. These results show that the ERCOT market had sufficient responsive reserve capability on the system in all hours.

Although not obvious from the scatter plot, prices are negatively correlated to the responsive reserves capability. Prices are expected to be low when there is a surplus since the marginal costs of supplying responsive reserves should be zero. However, the figure shows that prices are frequently more than $10/MWh when hundreds or thousands of megawatts of surplus responsive reserve capable resources are available. This is surprising because the marginal costs of providing responsive reserves from an online unit are generally extremely low. We can only attribute these results to the sequential nature of the operating reserves market, which indicate the potential benefits of jointly optimizing the operating reserves and energy markets. We would recommend joint optimization of these markets in the context of the nodal market designs currently under consideration in Texas.

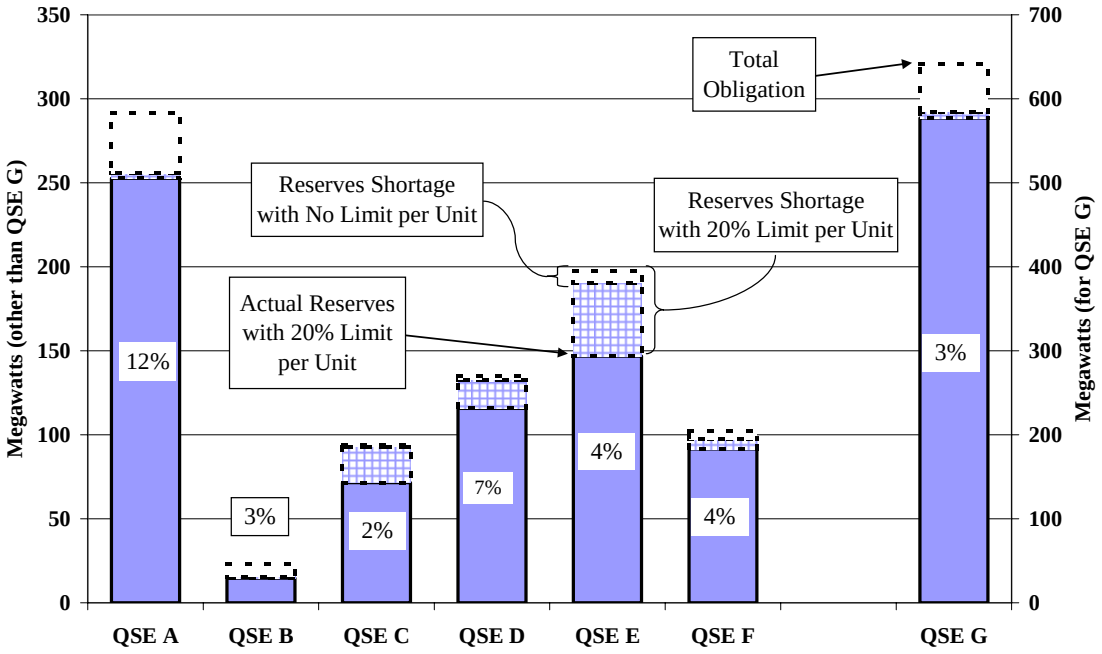
Although, there is usually a significant excess of responsive reserves in real-time, payments are made only to QSEs that scheduled to provide responsive reserves day ahead. These QSEs must ensure that sufficient capacity is on-line in real-time and not scheduled for energy or another ancillary service. Furthermore, QSEs cannot offer responsive reserves capacity into the balancing energy market. In the event that responsive reserves are deployed, QSEs must be capable of ramping up the reserved capacity within 10 minutes. ERCOT requires that the amount of responsive reserves that can be provided by any particular resource must be 20 percent or less of the capacity of the resource. This requirement is intended to ensure that responsive reserves are held on a diversified set of units. If one or two units providing responsive reserves trip off-line or are congested down, it will not greatly impact the reliability of the system.

Under normal circumstances, responsive reserves are not actually deployed in real-time. This makes it difficult to know whether QSEs with reserves obligations would have responded in the event that responsive reserves were deployed. However, it is possible to evaluate whether the QSE has the capability to respond. The following analysis evaluates whether QSEs set aside a sufficient quantity of un-loaded, ramp-capable, on-line capacity to meet serve their responsive reserves obligations in real-time. For most QSEs, the total capacity required to serve energy and ancillary services obligations is largest during the peak load hour. QSEs must commit sufficient resources to serve their obligations in the

peak hour and are likely to have excess capacity during other hours. This analysis focuses on the peak load hour of the day, when QSEs are most likely to have insufficient capacity to serve their obligations and when the system-wide responsive reserve capability is likely to be the lowest.

Figure 50 shows a summary of the analysis including only those QSEs that were short of their responsive reserves obligation during the peak load hour on more than 2 percent of the days during the study period. Shortages were detected by analyzing the ability of each QSE to ramp up their online resources within 10 minutes to satisfy their responsive reserve obligation, given the QSE's unit specific ramp rate limitations.

Figure 50: QSEs Failing to Satisfy their Responsive Reserve Obligations January to August 2004



The percentage shown on each bar in Figure 50 is the percentage of days when the QSE did not satisfy its reserve obligation in the peak hour. The quantities in the figure show the average reserve obligation and the amount being provided by the QSE in only those periods when the QSE is not satisfying its obligations. The quantities for QSE G are shown against the secondary y-axis on the right because they are substantially larger than the quantities for the other QSEs.

This assessment is based on the capacity, ramp rate, and actual real-time output of the resources in each QSE portfolio. In some cases, QSEs set aside resources to provide reserves that are subsequently called upon to provide OOM up energy. To avoid showing QSEs in this position as having failed to satisfy their obligations, the analysis uses the scheduled generation in the real-time resource plan for the resource instead of actual output in these cases. While the six individual QSEs are short of their reserves obligations in a non-trivial portion of the time, the shortages by these six QSEs account for less than 1 percent of the market-wide responsive reserves requirement.

There are considerable incentives for QSEs to use some of this capacity to provide energy, particularly when the balancing energy market clearing price is very high. Although ERCOT calculates unloaded capacity by QSE, it does not monitor responsive reserve capability by QSE, which depends on the ramp rates of the QSE's units, their output levels, and their eligibility to provide responsive reserves. To improve the performance of the QSEs in meeting their responsive reserve obligations, we recommend that ERCOT monitor the responsive reserves held in real-time and withhold responsive reserves and energy payments corresponding to the quantity of reserves that were not maintained. ERCOT or the PUCT may also want to impose penalties on QSEs who have failed to meet their obligations.

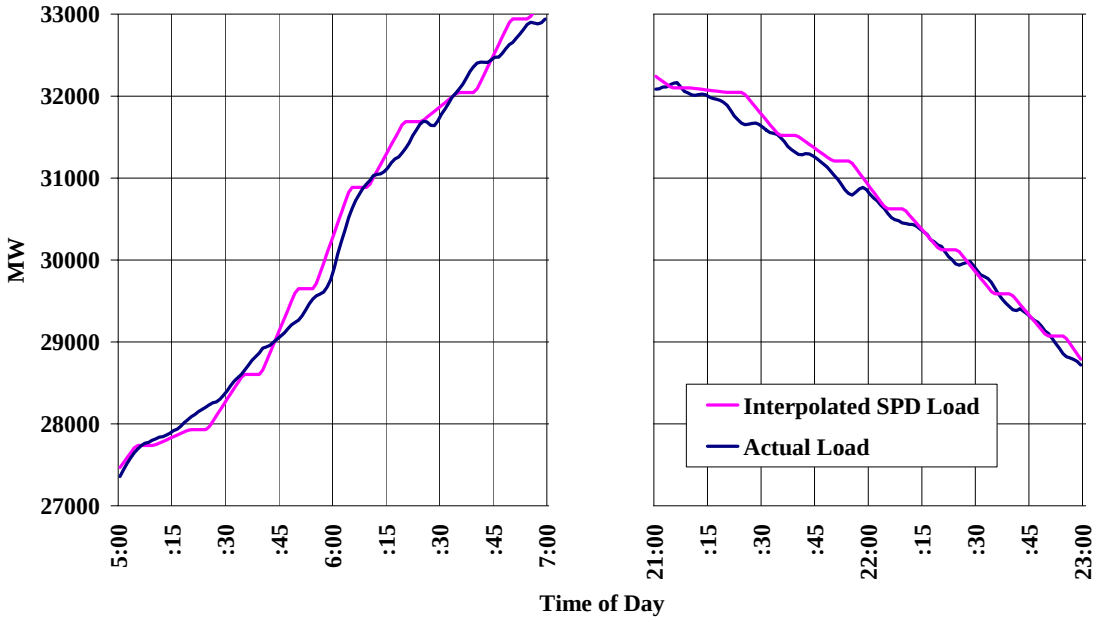
The QSE most frequently short of its responsive reserves obligation was QSE A, which was short on 12 percent of the days. On these days, QSE A's average obligation was 292 MW, while it maintained only 253 MW (an average shortage of 39 MW). The amount by which QSE A was short of capacity to provide responsive reserves was not very sensitive to ERCOT's 20 percent limitation. Disregarding the 20 percent per unit limitation, QSE A's average shortage of responsive reserves in these intervals reduces to 36 MW. The shortages of reserves observed in Figure 50 for the other QSEs are infrequent. To the extent that shortages occur, the amounts of shortage are highly sensitive to the 20 percent per unit limitation. For these QSEs, 86 percent to 94 percent of the shortages would not occur without the 20 percent limitation.

The 20 percent limitation is intended to ensure that responsive reserves are held on a number of different units and that those units will be likely to respond to a reserve deployment. However, for many relatively small units, the 20 percent per unit rule is overly restrictive and may contribute to inefficiently high responsive reserve costs. Hence, we recommend modifying the restriction to make it less binding on relatively small units. For example, the restriction could be the higher of 50 MW or 20 percent of the unit's capacity. This would effectively eliminate the restriction on units smaller than 250 MW while still ensuring that the responsive reserves would be held on many different units.

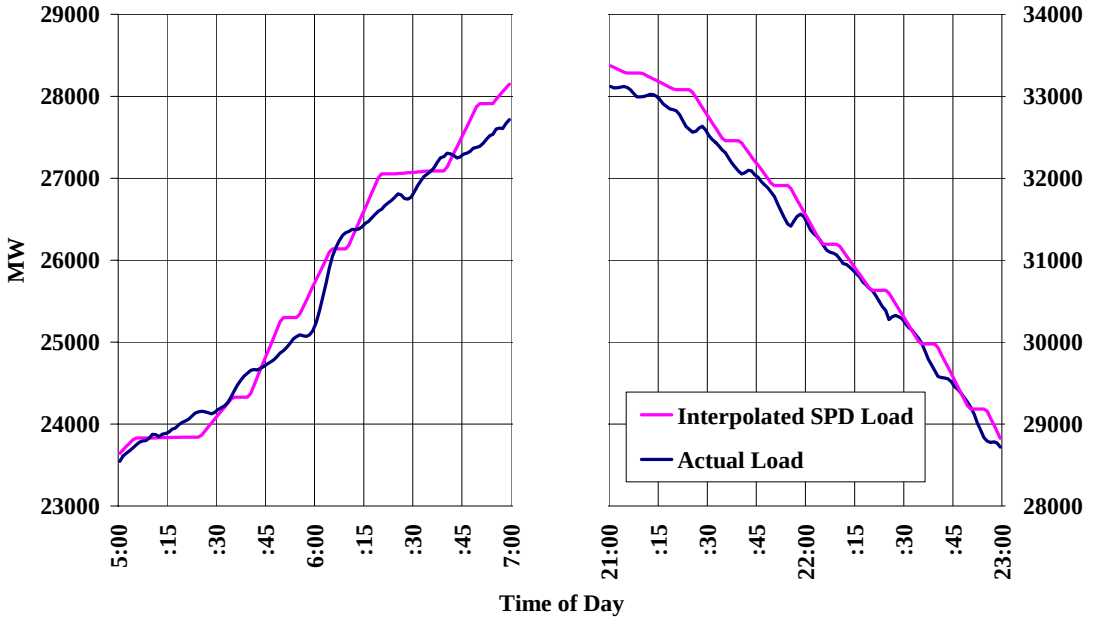
Although responsive reserves are procured on a portfolio basis, this recommendation is still important because it will allow many QSEs to supply more responsive reserves from their portfolios.

APPENDIX A

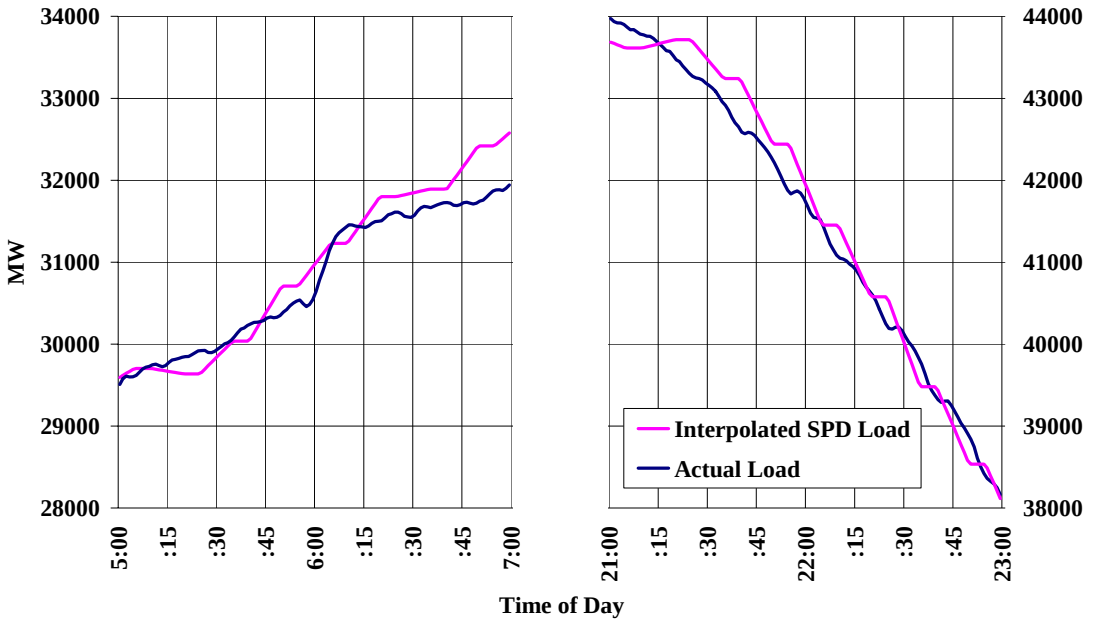
Schedule Control Error vs. Load Forecast Error During Day-Night Switch Hours (Jan - Feb 2004)



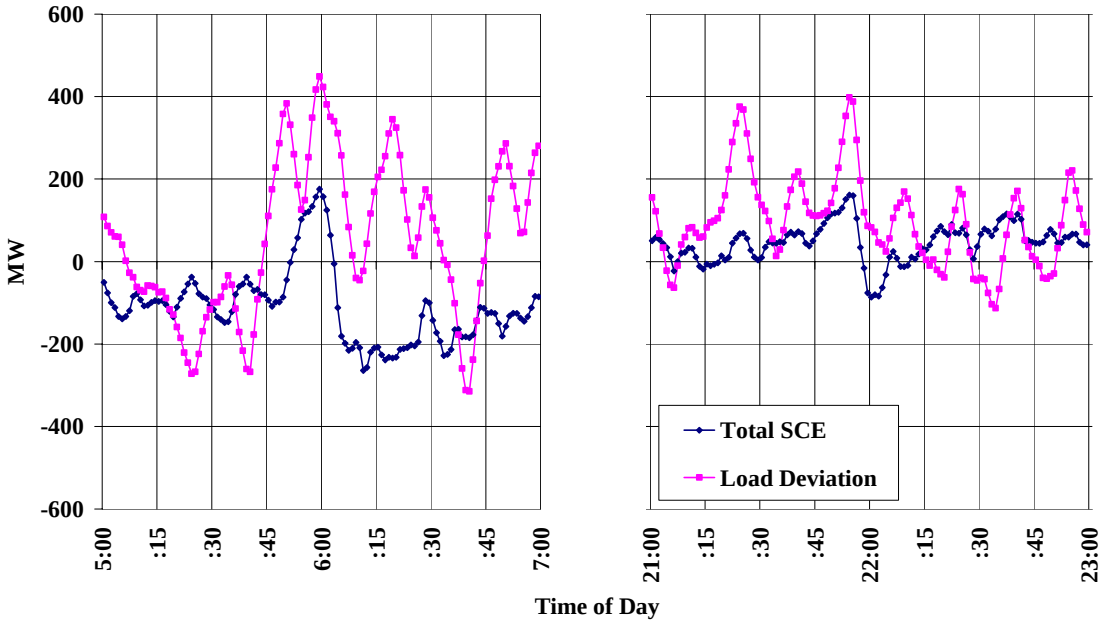
Schedule Control Error vs. Load Forecast Error During Day-Night Switch Hours (Mar - May 2004)



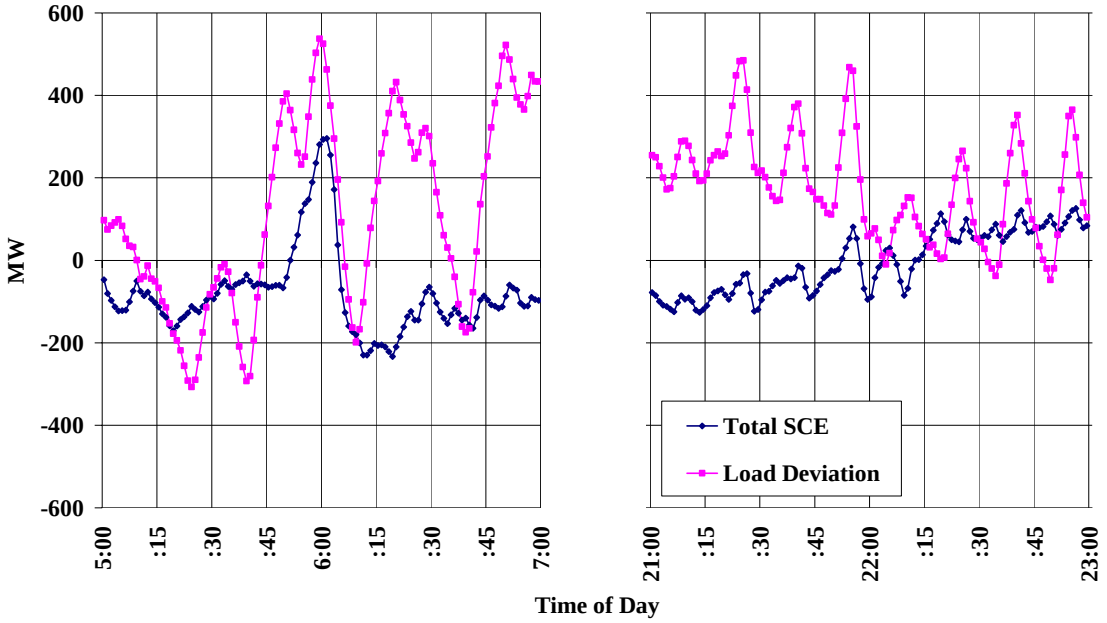
Schedule Control Error vs. Load Forecast Error During Day-Night Switch Hours (Jun - Aug 2004)



Schedule Control Error vs. Load Forecast Error During Day-Night Switch Hours (Jan - Feb 2004)



Schedule Control Error vs. Load Forecast Error During Day-Night Switch Hours (Mar - May 2004)



Schedule Control Error vs. Load Forecast Error During Day-Night Switch Hours (Jun - Aug 2004)