SPEER

The South-central Partnership for Energy Efficiency as a Resource

The Debate About Demand Response andWholesale Electricity Markets

Authors:
Robert King, CEO, SPEER
Jaden Crawford, Wholesale Electricity Markets Subject Matter Expert
Barry Huddleston, M.S.
Steve Isser, President, Energy Law & Economics, Inc.

October 2015

Contents

This paper was produced by SPEER, the South-central Partnership for Energy Efficiency as a Resource, under funding by the Heising-Simons Foundation. Contributors include Steve Isser, Ph.D., President, Energy Law & Economics, Inc., and Author of Electricity Restructuring in the United States: Markets and Policy from the 1978 Energy Act to the Present; Barry Huddleston, M.S., and Robert J. King, P.E., President of Good Company Associates, and SPEER's CEO.

Executive Summary

Demand response (DR) is the relatively recent term for the reduction or curtailment of a customer's electric consumption (demand) in response to market or utility signals of one form or another. Early market opportunities for customer participation allowed loads to respond to emergency conditions, to avoid potential outages. More recently this new resource has become viewed as a form of planned or operating reserve capacity to help increase reliability, reduce outages, avoid unnecessary capital investments, integrate intermittent resources, or even balance the frequency of the power on the grid.

Still more recently, demand response has been allowed to participate directly in energy markets, competing directly with generation to satisfy market demand, and contributing to price formation. This increasing competition from inclusion of demand resources is the next logical step for some more sophisticated customers and service providers as they and markets co-evolve. Allowing and incenting customers to help meet market-wide demand for power through strategic load reductions can help reduce short-term costs, avoid unnecessary capital investments in generation as well as transmission and distribution, and help mitigate price distortions associated with market power.

The Supreme Court recently heard a case related to the price of electricity in the US that could affect consumers in a number of ways. It will decide the authority of the Federal Energy Regulatory Commission (FERC) to regulate the participation of demand response in wholesale electric markets and how it is to be compensated. So this case will both affect benefits to customers that receive payments for their load contributions, and all customers in FERC jurisdictional markets who benefit from the larger savings created by this participation.

The Electric Reliability Council of Texas (ERCOT) is not subject to FERC regulations because it exists wholly with the boundaries of the state of Texas. Nevertheless, Texans are not immune to national debates on economic and legal matters that affect our market design. As case in point, ERCOT has allowed DR to participate in its wholesale market in a variety of ways. The following paper will discuss the evolution of wholesale markets generally, as well as the evolution of customer participation in those markets. We will also discuss the status of demand response in the ERCOT market specifically.

We find that daily energy markets are an appropriate and organic expansion of the opportunities for customer demand to be expressed productively and efficiently in the market. And, we find that potential consumer savings are substantial enough to make this a topic worthy of study. The analysis appended to this paper[1] indicates that wholesale market prices in ERCOT can be reduced substantially by the participation of even a modest increment of additional demand response in this energy market. We found an incremental demand response of 1500 MW or less, in a few critical hours, on only five days across the mild weather years of 2012 and 2013, could have led to a total reduction in spot market costs of as much as $200 million.

[1] Appendix A: Incremental Demand Response Analysis: An ERCOT Case Study, SPEER, May 2015

Economists have viewed the market for electricity as historically inefficient. The price of the product has generally been very stable, usually being set to represent an average unit cost of electricity over time because it has existed in a fully regulated environment. The advent of wholesale and then retail competition, where adopted, has allowed changes in wholesale prices driven by movement in fuel costs to be expressed more quickly in retail prices. So, for example, 2008 electric prices in Texas jumped quickly when natural gas delivery was impaired, in part by hurricane damage to coastal facilities, and then dropped again as production came back online. In addition, electric meters have traditionally provided only simple monthly consumption information, leaving customers little information or incentive for modifying their short-term use and electric providers little opportunity for time variant rates.

For these reasons, as well as the fact that energy using buildings and appliances have long-lifetimes and represent large investments, the market for electricity has traditionally been characterized by very inelastic demand. That is, customers are slow to alter their demand in response to changes in price. Innovations accompanying retail market competition, where it is has been adopted, include the large scale deployment of smart energy meters and the emergence of communicating digital technologies which are giving even smaller customers more granular information about their energy use, and more control over their energy appliances and demand patterns. The adoption of retail competition is often justified by its potential to make wholesale price variations more transparent to retail customers.

In fairness, however, while time of use rates may be the economists' remedy for bringing demand elasticity and efficiency to the electric market, voluntary adoption of price variant rates has been slower than many hoped. And few regulators have attempted to mandate time of use rates. Customers, who in the short run can easily control only a fraction of their loads, will not be attracted to time varying prices. A customer that can only shift 30% of its load in response to a daily price swing will not likely benefit from a time of use rate that penalizes the 70% remaining unchanged. This has in part been responsible for the development of demand response products and markets—in which customers can purchase electricity at retail on a flat rate, based on average costs—and receive a price signal in the form of a positive incentive for curtailing the load that can be reduced with short notice.

Perhaps the most important innovation recently has been the emergence of a class of customer energy management services, provided by Load Serving Entities (LSEs) or independent intermediary companies (sometimes in partnership) that have the sophistication to track the real-time volatility of wholesale markets and to enable customers to respond to market signals with short-term alterations in demand. In most cases, the customer response is automated using sophisticated technological systems that adjust equipment to price signals with no action taken by the customer. Markets have simultaneously evolved opportunities for these loads, or aggregations of these loads, to contribute to the efficiency of the larger electric markets. It is really the work of these service providers whose effort and expertise is enticing more customers to participate in market opportunities that have turned customer loads into a more immediate and effective energy resource for the wholesale markets.

With respect to the appropriate compensation of customer participation, the national debate centers around whether customer load curtailments should be paid exactly the same as generation

offered into a market to meet demand, which is what FERC determined should be done. Generators pointed out, and the DC Circuit Court agreed, that this amounts to an unequal payment to loads, because the loads also experience a reduction in energy costs. That is, the appropriate compensation should be the locational marginal price paid to generation (LMP), less the customer's actual retail cost of generation for the energy not consumed (LMP-G). This is the informal policy under which ERCOT currently operates. This pricing ignores the potential opportunity cost of a customer's curtailment, or real operational costs associated with participation in the market for the load, much less for the intermediary curtailment service providers that have emerged to create this efficiency. Nevertheless, we accepted the premise for review, but after struggling with this issue ourselves, we find that the cost and complexity of implementing this is likely to outweigh the benefits of implementing something workable if less precise.

When evaluating the LMP versus LMP-G debate, we rely again on our own case study evaluation of an incremental increase in demand response in ERCOT during 2012 and 2013 referenced above. We find that the differential in cost for paying loads full LMP rather than the more conservative LMP-G is likely less than one percent of the total potential price savings to the spot market, or less than $2 million for savings of as much as $200 million. To actually track what a customer would have paid for what it would have used in the absence of its curtailment, and to assure that its retail load serving entity and its energy curtailment service provider are not disadvantaged unfairly, turns out to be nearly impossible to do accurately. Administrative and market procedures to implement LMP-G are bound to lead to defining similarly imperfect, if workable solutions that attract less consumer participation and reduce potentially significant benefits. We are led to conclude that compensating loads at full LMP may actually be the most efficient solution.

Oral arguments were heard on these topics before the Supreme Court in mid-October. Whatever the Supreme Court decides, there are demonstrable benefits to all customers associated with demand participation in electric markets, and ERCOT is not bound by the Supreme Court's decision on FERC. ERCOT stakeholders have struggled to define a workable LMP-G solution for over four years and continue to struggle. The cost difference of full LMP is so slight and the benefits of demand response so large, it would seem that an effort to create a likely unattainable perfect solution is preventing the attainment of hundreds of millions of dollars of savings to the market.

The full paper dives much deeper into the underpinnings and history of this debate, in the hopes that it will bring some much needed clarity to this important issue. Added as an appendix is the calculation of the actual impact DR would have had if it could have more fully participated in the ERCOT market. That impact is substantial: over $200 million saved over five days in 2012-13. A market that sends a price signal to reduce peak consumption will lead to lower prices. Placing barriers to load participation has—and will continue to have—the opposite effect.

I. INTRODUCTION

The Supreme Court has heard the appeal of the D.C. Appeals Court judgment about a fundamental element of the evolving wholesale markets for electricity in this country. The decision to be reviewed involves the Federal Energy Regulatory Commission's (FERC) direction to interstate markets on the inclusion of, and payment for, demand response in auctions that set the price of electric power in wholesale markets.[2] Demand response (DR), the voluntary reduction of energy use, can potentially make a significant contribution to price moderation in wholesale electricity markets. In independent research, we found that a modest amount of incremental demand response would have reduced the short-term price of wholesale power in Texas substantially.[3] Even in the mild weather years of 2012 and 2013, an additional 1500 MW of demand response could have lowered wholesale power costs nearly $200 million in just a few hours over only five days. So, while ERCOT, the Electric Reliability Council of Texas, exists wholly within Texas and is, therefore, not subject to the wholesale market regulations of FERC,[4] we are not unaffected by the logic of national legal and economic debate, and this case is still of some interest in ERCOT.

When the Federal Energy Regulatory Commission (FERC) proposed to compensate demand response resources in wholesale markets at the full market price of energy,[5] it triggered an intense discussion among market participants and their consultants. This heated conversation continued through a year of filings and a technical conference, and spilled over into several editions of the *Electricity Journal* during that period. The FERC adopted this proposal in Order 745.[6]

---

[2] *Demand Response Compensation in Organized Wholesale Energy Markets*, Order No. 745, FERC Stats. & Regs. ¶ 31,322 as codified,18 CFR Part 35. We refer to this henceforth as simply FERC Order 745.

[3] Incremental Demand Response Analysis, SPEER, May 2015

[4] Federal regulation of wholesale markets by FERC, the Federal Energy Regulatory Commission, is predicated upon the authority granted FERC by the Federal Power Act over interstate commerce in electricity. The ERCOT grid is an isolated electric transmission grid with only High Voltage Direct Current connections to out of state transmission grids, serving near 80% of the Texas market. See Jared Fleisher, ERCOT's Jurisdictional Status: A Legal History and Contemporary Appraisal, 3 Texas Journal of Oil, Gas, and Energy Law 5 5 (2008) FERC does have indirect authority over ERCOT through its oversight of NERC, the Energy Reliability Organization, and its regulation of reliability under the Energy Policy Act of 2005.

[5] Demand Response Compensation in Organized Wholesale Energy Markets, Notice of Proposed Rulemaking, 75 FR 15362 (Mar. 29, 2010), FERC Stats. & Regs. ¶ 32,656 (2010) (NOPR).

[6] *Demand Response Compensation in Organized Wholesale Energy Markets*, Order No. 745, FERC Stats. & Regs. ¶ 31,322, *order on reh'g & clarification*, Order No. 745-A, 137 FERC ¶ 61,215 (2011), *reh'g denied*, Order No. 745-B, 138 FERC ¶ 61,148 (2012), *vacated*, *Elec. Power Supply Ass'n v. FERC*, 753 F.3d 216 (D.C. Cir. 2014) (*EPSA*), *cert. granted*, 83 U.S.L.W. 3835 (U.S. May 4, 2015) (No. 14-840) *& consolidated sub nom. EnerNOC, Inc. v. Elec. Power Supply Ass'n*, 83 U.S.L.W. 3835 (U.S. May 4, 2015) (No. 14-841).

EPSA, the Electric Power Supply Association, and other interveners challenged the FERC decision before the DC Circuit Court of Appeals, essentially claiming that a customer must first pay for energy to offer to sell it back into the market, and that in any event this is simply a customer's decision to forgo a retail purchase. A panel of judges agreed that paying customers the full market price not to use energy creates an inappropriate subsidy, and violates the mandate that rates be "just and reasonable,"[7] and went further to say that demand response is a retail transaction, and therefore outside the wholesale market authority of FERC. The decision of the DC Circuit to vacate Order 745 revolved around these two issues: whether FERC's rule entails direct regulation of the retail market—a matter exclusively within state control—and thus exceeds the Commission's authority,[8] and whether FERC adopting market pricing for demand response resources was "arbitrary and capricious" because Order 745 would result in unjust and discriminatory rates.[9] This DC Circuit decision was appealed to the Supreme Court by the US Solicitor General and the Court heard the case earlier this October.

Although legally unaffected by the outcome of the federal ruling, this case could indirectly influence regulation of demand response in the ERCOT energy market. As will be described herein, Texas has determined that it is appropriate and efficient to include the participation of demand response in ERCOT's wholesale electric market in a variety of ways, with good reason. Whether the mandate to pay the full energy market price to demand response resources was just and reasonable, on the other hand, will be examined here further, because this issue is one that has yet to be formally resolved in the ERCOT market as well.[10] Although Texas is today operating according to the more conservative pricing position represented by EPSA, strict adherence to this principle has its own costs. Interestingly, our recent analysis[11] suggests that paying the full energy price for the substantial potential consumer benefit, mentioned above ($200 million), would entail an added cost of less than $1 million over the EPSA position on pricing. Creating the settlement system to avoid this additional payment for those loads contributing to the market could incur other costs and create barriers to participation by demand response resources. The danger seems to be that --as Pat Wood, a former chairman of both the PUC of Texas and FERC, used to say--"we may let the perfect become the enemy of the good."

7 See Steve Isser, Just and Reasonable: The Cornerstone of Energy Regulation (June 30, 2015). Available at SSRN: http://ssrn.com/abstract=2625131 or http://dx.doi.org/10.2139/ssrn.2625131, for a discussion of the application of the term "just and reasonable" in electricity regulation.

8 EPSA, 753 F.3d at 224.

9 *Id*. At 225.

10 We will refrain from opining on the legal issues concerning the validity of the Court's judgment. This issue has been thoroughly vetted by eminent legal scholars. See Brief for Energy Law Scholars, as Amicus Curie Supporting Petitioners, FERC v EPSA (Nos. 14-840, 14-841) (July 16, 2015).

11 Incremental Demand Response Analysis, SPEER, May, 2015

II. WHOLESALE ELECTRICITY MARKETS

A. *Electricity Markets in a Nutshell*

Prior to wholesale competition in power generation, electricity was primarily generated and transmitted by vertically integrated electricity companies, most of which were investor-owned utilities. They owned the generation plants and transmission and distribution wires that carried power from the power plants to the customer. While some of these utilities were owned by holding companies, and many participated in power pools, most power was generated and sold in-state. The price of electricity was established by state regulatory commissions, based on cost of service considerations.

With the advent of electricity restructuring, in the wake of the Energy Policy Act of 1992,[12] encouraged by FERC[13], centralized electricity markets were established in numerous regions, and now encompass two-thirds of the electricity sold at the wholesale level in the United States. These markets were based on historical power pools, such as the New England Power Pool which became the New England Independent System Operator (NE-ISO), the New York ISO (NYISO), the Pennsylvania-NJ-Maryland interchange (PJM), the Midwest Power Pool (MISO), and the Electric Reliability Council of Texas (ERCOT). The California ISO (CAISO) and the Southwest Power Pool (SPP) were created de novo.

These centralized markets developed complex software programs to efficiently match supply to demand, taking into account transmission constraints, ramping constraints (the speed at which a generator can increase or decrease output) and other issues that impact instantaneous power quality and system reliability. The unit dispatch software monitors and evaluates system security and performance and dispatches generation units. Unit dispatch software was designed to minimize generation costs subject to reliability constraints, thus the term, security constrained economic dispatch (SCED) is applied to the operating software behind our modern energy markets. Control centers must ensure that generating units will be ready when needed to follow the daily load cycle that the transmission system is capable of carrying the loads, and that backup generating capacity is available in case of equipment failure.[14]

Many wholesale markets have also evolved a separate "capacity market," through which system operators hold an auction to acquire the commitment of generators to make a certain level of power capacity available at designated future times. This can be viewed as an insurance program to assure that sufficient capacity investments are made to meet projected load growth. In these markets, where some component of generation is thereby supported by a capacity payment, energy payments will reflect a larger supply of generation supported by capacity payments. Interestingly, ERCOT is an "energy-only" market, meaning that generators must recoup all their costs and earn profits based upon the energy prices they receive, and the other, or "ancillary" products they can sell the grid operator.

---

[12] Pub. L. No. 102-486, 106 Stat. 2776 (1992), codified at, among other places, 15 U.S.C. § 79z-5a and 16 U.S.C. §§ 796(22-25), 824j-l.

[13] FERC Orders 888, 889, 890 and 2000

[14] Steve Isser, *Electricity Restructuring in the United States: Markets and Policy from the 1978 Energy Act to the Present* (Cambridge: Cambridge University Press): 124.

System operators attempt to maintain system frequency at 60 hertz using ancillary services to balance generation and load. The key ancillary services are Regulation, Spinning Reserves, and Non-spinning Reserves. Traditionally, regulation is the use of online generating units equipped with governors and automatic generation control that can change output quickly in response to control authority signals. Spinning Reserve is the use of generating equipment that is online and synchronized with the grid that can begin to increase output immediately and be fully available within ten minutes. Non-spinning Reserve is comprised of generating equipment that can be synchronized with the grid, usually within thirty minutes.[15]

Each existing control authority operates its electric market(s)[16] using SCED models to compare multiple offers for energy generation and these added support services, and bids to buy them, and arrive at an efficient solution in each interval of time, given the need to respect transmission limits and other reliability considerations. Grid operators then use software to issue instructions, usually every five minutes, directing generators to increase or decrease output (dispatch instruction), taking into account ramping constraints (the rate at which a generator can change output). The market outcomes generated by this software are "workably efficient" that is, given the complexity of the software and various technical considerations such as the nonlinear nature of some constraints, the outcomes are about as efficient as can be reasonably achieved.

The inefficiency of electric markets, as Robert Borlick has pointed out, is based on the absence of retail tariffs that directly and dynamically reflect the wholesale spot market prices to customers.[17] The adoption of retail competition in various regions of the country was undertaken at least in part with the hope that these markets would develop this price transparency. Historically, however, short-term demand for electricity has been relatively inelastic. That is, in the long run most consumers will respond in some manner—develop more efficient processes, move into more efficient homes, purchase more efficient replacement appliances—but in the short term, customers have lacked good information about how their energy use during the day was effecting their bills. And, what feedback they received was very delayed, often by at least 30 to 60 days.

Of course some price variation is somewhat predictable. Traditionally, it was assumed that cost increased as peak load increased and more expensive generation units were employed. Load tends to peak in the afternoon in the summer (early morning in the winter months), and the highest peaks are during extreme temperature events. As demand approaches the limits of supply, each added increment of demand imposes increasingly more costs than the previous increment, as the most expensive generators were deployed (peaking units that only ran a few

[15] *Id*. at 127-28.

[16] Most wholesale electricity markets actually operate two interlinked markets, a day-ahead market which establish financial commitments to buy and sell power, and a real-time market in which physical transactions occur and the control operator attempts to balance physical supply and demand at all times. Similar software is used to dispatch generators in both markets, but there is usually some divergence between the markets due to unexpected contingencies (outages by generators or failures of transmission lines, changes in actual load due to unexpected weather events, etc.). The day-ahead market acts to mitigate the exercise of market power in the real-time market by allowing market participants to hedge against potential price spikes in real-time.

[17] Comments of Robert L. Borlick, Energy Consultant, Demand Response Compensation in Organized Markets, Docket No. RM10-17-000, May 13, 2010.

hours per year). Time-of-use rates are a means to send customers price signals in line with predictable price variations such as these. And the expanding adoption of automated meter-data infrastructure is improving the capacity of the system to support improved response to market signals. Despite encouraging technology development, most buildings do not yet have the intelligence to respond to these signals automatically, however, and time of use rates may not be optimal for the market and the customer in every circumstance in any event. For example, a customer than can easily shift 30 percent of its load off peak, may still be penalized from picking a time of use price differential that penalizes the 70 percent of its load which will still remain on peak.

Prices can also spike during "off-peak" periods, too, because a market operator can only dispatch those resources that are ready to operate. If there is an unexpected outage or demand peak, units that weren't chosen earlier, are turned off for necessary maintenance, or have dedicated their output to providing ancillary services, may not be available to the energy market, at least without a significant delay. Unless customers are enabled to respond by reducing demand at these times, prices may spike for an hour or two while off-line resources ramp up. Price fluctuations have been exacerbated by increased market penetration of renewable resources that have variable outputs (a cloud passing over a solar farm or a sudden drop in wind). Over a year, these short-term price spikes can result in significant increased costs for consumers.

So, in the long run, all customers may respond to price signals in some manner, and an increasing number may be drawn into changing their behavior in general ways, but, only by creating a pathway for customers to respond to market price volatility, as it happens, in real time, can demand interact with supply in the process of efficiently determining prices. Perhaps the greatest gift of competitive markets in this regard has been the appearance of entrepreneurial intermediaries, in the form of energy curtailment service providers. These companies and progressive load serving entities (sometimes in partnership), are providing customers with the technology, information and support they need to become active participants in the wholesale energy markets at a more granular level than before. And, these new entrants, working with market operators and other stakeholders, are helping to evolve avenues through which this potential efficiency can be practically manifested.

B. *What Is Demand Response?*

Demand Response (DR) is the process of customers committing to reduce demand in reaction to high energy prices, demand charges, or other incentives or signals from an energy provider, utility or grid operator to support system reliability. Demand response resources must have a degree of predictability and reliability to qualify for participation in markets as such, or be liable to penalties or disqualification. Price responsive demand is not a wholesale market demand response resource *per se*. Rather, this refers to customers who face retail electricity prices that vary with the market energy price adjust their consumption as price changes or reaches a threshold, a decision independent of the operation of electricity markets. This would rightfully be considered a retail decision not to consume a product. Smart meters allow access to the data required to bill customers at the market price, and allow third parties access to customer information needed to aid them in controlling their loads. Experience has shown, however, that customers will adopt time sensitive rates less slowly than market advocates may have hoped.

And, demand response service providers find customers respond much more readily to offerings which are more structured and less uncertain than a simple price response strategy.

The term "Demand Response" grew out of the evolution of market mechanisms to assist and incent customers, or third-party managed aggregations of customer loads, offer a quantifiable and predictable resource directly in wholesale markets. Initially, the primary markets for demand response resources have been emergency demand programs established by the Independent System Operators, or ISOs, that administer wholesale electric markets. These programs usually operate by requiring participants to decrease load on the command of the ISO control operator, usually within an hour of notification or less, for as long as instructed by the ISO, up to some program limit. There is often also a limit to how often these resources can be called in a year.

These resources are paid on the basis of capacity, the MW of potential load shed participating in the program, and sometimes may be compensated with an additional energy payment for the period of time these resources are actually called upon to reduce energy consumption. They are usually called as an emergency measure just before the ISO runs out of generation capacity, and before involuntary load shedding (rotating blackouts) commences.

With the establishment of formal capacity markets, some of these resources in these areas migrated to the capacity markets, which permitted demand response resources, and in some instances, energy efficiency, to bid in as capacity in these markets. The requirements of the capacity markets tend to be similar to those of the emergency demand response programs.

Controllable loads, like motors with variable speed drives, with the ability to alter their level of consumption quickly in response to market, or market operator signals, have been allowed to provide more ancillary services with more exacting performance requirements, and activities in this area is growing. Regulation Service, for example, is the "fine tuning" knob of control, as we discussed in the previous section. The demand response resources capable of participating in this market tend to be more responsive than generation, making them higher quality resources for this use.

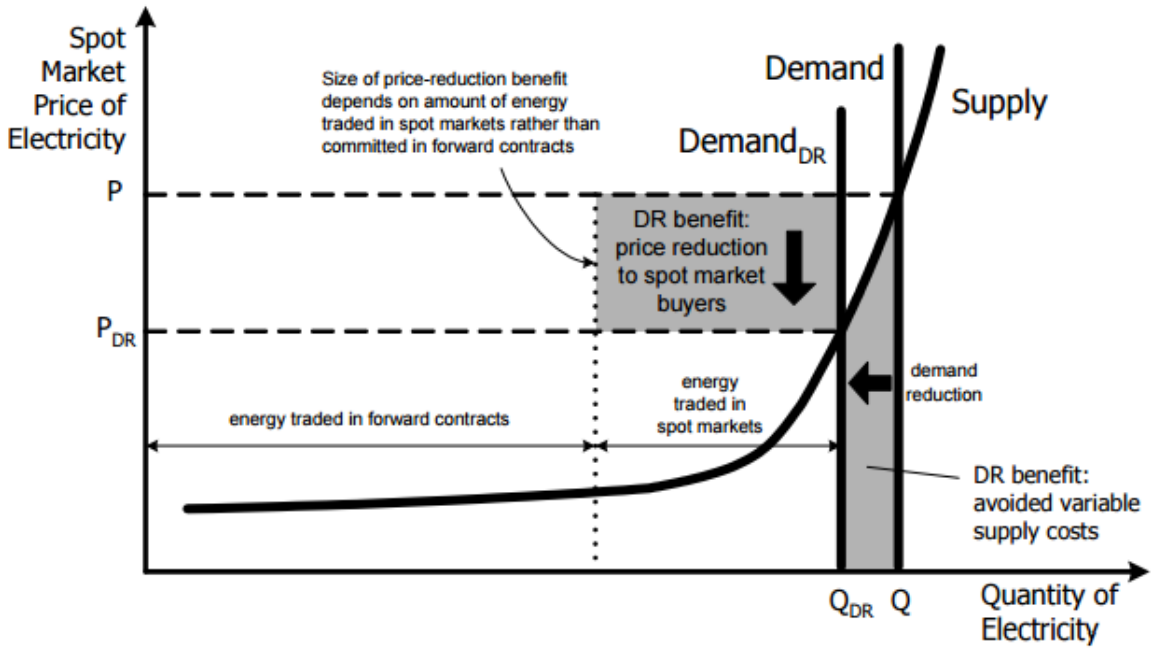
Spinning reserves present similar demands, though usually not as stringent in terms of speed of response. In the case of supplemental or Non-spinning Reserve Service, generally purchased by markets as insurance for periods of larger unexpected swings in demand, there is a wider range of requirements between markets, but certain loads can compete with generators to provide these service markets.

Demand response resource participation in wholesale electricity markets is a more recent innovation. Resources place price bids, and if their bid is accepted, must be able to drop or increase load when instructed. Such economic demand response can potentially provide substantial benefits by increasing the supply of resources that can balance supply and demand in near real time. The magnitude of possible benefits depends on how widely average and marginal costs diverge, and how frequently those costs diverge by a substantial margin. In a highly constrained market, when peak demand can be very close to supply limits, the potential short-term efficiency benefit from this kind of demand response program might be quite substantial, as

suggested by our own analysis of the incremental impact of additional demand response in the ERCOT market during the relatively mild weather years of 2012 and 2013.[18]

Figure 1 depicts graphically how demand reductions can alter marginal clearing prices in wholesale markets.[19]

Figure 1: *Impact of Demand Response in Regions with Organized Wholesale Markets*



As evidenced here, the supply curve is developed by "stacking" offers from generating and demand-side resources in merit order from lowest to highest. Because of the intense competition between and among resource suppliers, offers largely reflect marginal operating and maintenance costs (and, occasionally, may also reflect recovery of some or all fixed costs in energy-only markets) of generating resources, and a "strike price" for economic DR. Load serving entities such as utilities or competitive REPs[20] also bid to buy energy to meet expected load,, or demand.

[18] ERCOT made real-time market interval data available for these two years with respect to the actual offers of generation accepted by the market. Data from the unusually extreme year 2011 was for some reason not available, but would have lead to much more striking results.

[19] U.S. Department of Energy (February 2006). *A Report to the United States Congress Pursuant to Section 1252 of the Energy Policy Act of 2005.* p. 73

[20] In ERCOT, the state legislature approved the creation of a wholesale power market in 1997. All load serving entities and generators participate in this market. Later in 1999, the legislature opened the service areas of investor-owned utilities directly regulated by the state to retail competition, unbundling the competitive functions of generation and retail sales from the still regulated utility functions including poles and wires and meters. So the remaining customer-owned cooperative and municipal utilities are allowed to remain vertically integrated under the regulation of their boards or city councils respectively, and serve their customer loads as they have traditionally, unless they opt into competition. In the investor-owned regions, the wires utilities cannot own or generate power, and customers are served by a competitive retailer of their choice. These retailers arrange for the acquisition and transmission and distribution of a sufficient amount of electricity to serve their customers load.

The market-clearing price is, thus, determined by the intersection of these curves, at the point where the offer price of the last supply resource that must be dispatched to serve the loads meets the demand.[21] With voluntary load reductions integrated into this market, the Demand curve would shift to $Demand_{DR}$, intersecting with the Supply resource curve at a lower point on the curve, and reducing the quantity demanded by load serving entities from Q to Q_{DR}; and, the market-clearing price from P to P_{DR}. Avoided variable supply costs are represented by the shaded portion under the Supply curve bounded by Q and Q_{DR}.

As our earlier study indicated, the reduction of market price resulting from a modest amount of incremental demand response, for only a few hours on a total of five days over two mild weather years could be substantial, assuming all energy is traded in real time. The actual benefit resulting from demand response driven price reductions depends not merely on the quantities demanded at the different prices in real time (the "spot market"), but also on the amount of energy traded outside the market through by-lateral ("forward") contracts between generators and load-serving entities. Since load-serving entities use a mix of long-term forward contracts and spot market purchases, the amount of customer savings in the wholesale spot market would depend on the price and amount of energy purchased in the respective markets. Future contract prices will be impacted by spot market prices, as the alternative to contracting is buying in the spot market, and expectations of future spot prices are influenced by current prices. If substantial load reductions routinely occur during periods of high demand and high prices, additional long-term savings might also result due to reductions in risk premiums in forward contracts. That is, if markets experience lower clearing prices and less volatility, suppliers may be willing to reflect the lower levels of risk in forward contract offers to load serving entities. This benefit of demand response participation in wholesale markets is harder to quantify.

III. DEMAND RESPONSE AND THE ERCOT MARKET

A. *Non-dispatchable Demand Response*

Some form of price responsiveness or non-dispatchable demand response, occurs in ERCOT as in other wholesale markets, and while not controlled by market operators, can nonetheless impact market outcomes over time. Utility-managed Load Management programs are designed to reduce peak demand to reduce the need for expansion of the transmission and distribution system or support the system operator. In ERCOT non-dispatched responses by customers includes attempts to reduce demand on the Four Coincident Peak days used to assign transmission charges to load. June, July, August and September are the peak generating months in Texas. The ERCOT system peak (the single point in time when the most generating capacity is needed) is set sometime during this four-month period. Load in each month will set a monthly

[21] Centrally dispatched, organized wholesale electric markets are operated using SCED engines. SCED optimizes energy and ancillary services (A/S) offers from the same pool of resources based on actual system conditions in real-time. Therefore, generating units might clear the market, and be dispatched, out of their energy offer merit order. This occurs due to the potential that other, lower-priced, infra-marginal MW are not cleared in the energy market but are instead used for various A/S, or congestion management. The co-optimization of energy and A/S reduces the overall cost of serving the market, even if the cost of energy production alone might be slightly higher than would otherwise be the case.

demand peak. If a Customer's end use load is running at the same time that ERCOT sets its monthly demand peak, then it will have contributed to the monthly coincident peak.

It can also include indirect response to wholesale market energy prices by a customer contracted to do so by its retail electric provider, or REP. Despite all these avenues through which customers may respond to price, however, a small minority of customers choose these paths voluntarily. A recent survey by ERCOT shows nearly 300,000 people signing up for TOU rates, including the innovative "free nights or weekends" rates introduced by competitive retailers.

B. *Load Resources in the Wholesale Market*

Load participation in the ERCOT wholesale market began with adoption of protocols allowing large industrial loads to offer Responsive Reserve Service (RRS) purchased by the market to protect against frequency excursions. Historically, utilities provided "interruptible service" tariffs to large industrial customers. In return for the right to interrupt a customer, the utility offered interruptible service customers a discounted rate, usually a reduction in demand charges. To pave the way for retail choice, all tariffs (including interruptible rates) offered by IOUs in ERCOT were terminated. Faced with the loss of this emergency cushion, and the loss of this benefit by industrial users, all stakeholders agreed to develop protocols for participation of these Load Resources (LRs)[22] as a spinning reserve, called Responsive Reserve Service. Requirements for LRs in RRS include at least 1 MW of interruptible load, with the telemetry to manually respond to a command to drop load within ten minutes, and an under-frequency relay that will immediately drop load if system frequency falls below 59.7 Hz locally. In 2003, LRs became eligible to supply 50 percent of responsive reserves.[23]

When LRs were adopted, the Texas market had a peak demand of around 40,000 MW, far smaller than the Eastern or Western interchanges. These interchanges could depend on the inertia of hundreds of generators to buy time to react to unexpected contingencies. In Texas, with the isolation from the Eastern and Western Interconnections, the outage of one large generator such as the South Texas Nuclear Project could potentially overwhelm the entire system. LRs provide a safety net, because of the automatic trigger (most areas used 59.0 Hz as the threshold for involuntary shedding of firm load, but LRs would be shed well before that point was reached). So while LRs fail to match the operational characteristics of responsive reserves provided by generation in most markets, their value as a safety net allowed the industrials to garner support from other participants in the ERCOT market.

More recently, under direction from the state's PUC, ERCOT created a second opportunity for load resources to help respond to emergency conditions. Emergency Response Service (ERS) is an alternative market for participation of LRs, including smaller loads or aggregations of loads as small as 100kW. ERCOT procures ERS to maintain grid stability during emergency conditions and reduce the likelihood of the need for rotating outages. ERS participants may offer to provide

[22] Initially called Load Acting as Resource (LaaR), and subsequently shortened to Load Resource or LR.

[23] ERCOT Board of Directors, Minutes of December 8, 2001 Meeting; ERCOT Technical Advisory Committee, Minutes, November 8, 2001 Meeting; *2004 ERCOT Methodologies for Determining Ancillary Service Requirements*, Proposal for Board Approval (November 18, 2003).

demand response with either a 10-minute (similar to LRs) or 30-minute ramp-up period requirement. ERS does not have the same telemetry and under-frequency relay requirements as RRS. ERS is defined by the PUC as a special emergency service (technically not an Ancillary Service). Rather than being purchased on a daily basis throughout the year, it is acquired in auctions three times a year for four months at a time, although there is some consideration being given to its conversion into a more regular daily market product. Additional provisions have also been made so that during the summer contract term (June through September), weather-sensitive ERS Load may participate under modified rules designed to encourage the reduction of air conditioning load.[24] All ERS resources are paid an availability or capacity payment if selected, whether called or not, although they must perform during random tests or actual events when they do occur to avoid payment reductions. There have been years when these resources are called upon multiple times (2011) or not at all (2012).

C. *Demand Response and Other Ancillary Markets*

Customers with demand response capability that can meet more exacting performance requirements can be qualified to provide Ancillary Services as Load Resources. A Controllable Load Resource (CLR), for example, is a Load Resource capable of controllably reducing or increasing consumption under dispatch orders (similar to automatic generation control). A CLR, such as a pump or motor with a variable speed drive, can provide Regulation, Responsive Reserves, and Non-spinning Reserves. Providers of these operating reserve services as well as ERS are eligible for capacity payments, regardless of whether the Resource is actually deployed.[25] It is interesting to note that, in the eligible Ancillary service markets, a Load Resource is compensated equal to the value of a generator available to increase its generation.

D. *Demand Response and the Energy Market*

With the support of their Load Serving Entity, Controllable Load Resources (CLRs) may enter bids into the ERCOT market auction, which is run by a SCED system underlying every transaction. SCED runs are executed every five minutes, so CLRs in SCED with bids at the localized marginal price (LMP)[26] must be capable of moving load incrementally in either direction every five minutes, based on SCED dispatch instructions. CLRs providing Non-Spin or Responsive Reserves are dispatched by SCED also, after their ancillary service capacity is released to the real-time market (not needed to provide energy). Aggregate Load Resources (ALRs) composed of multiple sites within a single ERCOT Load Zone may participate as a CLR.

---

[24] ERCOT Staff and DSWG Stakeholders, *Load Participation in the ERCOT Nodal Market*, Version 3.01, 2015, p. 10. This avenue was developed in recognition on the one hand that residential and commercial weather sensitive loads, such as air conditioning in particular, drive half or more of peak demand, and on the other hand, that these loads are only present to be shed during higher temperature times of the day.

[25] *Id*. at 10-12

[26] SCED programs calculate prices at each grid node, a point where power is injected into or taken from transmission lines, on the grid. These prices account for the cost of transmission congestion on the grid. When there is no congestion, prices are identical at every node, but when transmission is constrained, the congestion cost is the difference between the price at different nodes.

If a Load Resource's bid is on the margin, instructions could require the CLR to move its load up or down incrementally every 5 minutes.

So far there is very little load participation in the ERCOT SCED, because of these rigorous demands. ERCOT stakeholders are currently discussing a variety of protocol changes that could help to open up this participation, however, including a change that would allow all resources to bid in and be selected for multiple 5 minute intervals (rather than only one at a time), and permit longer response (ramp) times. While we have shown there are real benefits from loads participating directly in the process of wholesale price formation, these changes are probably necessary for these benefits to be realized.

It is also important to note that the current market is harmed by not allowing for the innovation that might come with a fully open and competitive market. ERCOT stakeholders' determination to deviate from the FERC recommended approach to compensation of load participants, combined with the unique structure of the Texas electricity market, presents significant barriers to third-party entrance into its real-time energy market. As noted above, Load Resources can only participate in the energy market today through their own Retail Electric Provider of Record or a vertically integrated utility. This limitation is present because if the market pays third parties for demand response at the same level as it does generation, the market operator will in turn collect less revenue for energy consumed than the total cost paid out. The uplift and fair allocation of this "cost" of demand response, presents significant administrative challenges and not-insignificant costs of its own when third parties are allowed to compete. Regardless of the fact that demand response may create total market savings far in excess of this gap in revenues collected by the market, who pays for this cost may be of primary interest to a market participant. One must ask, is the difference in compensation being debated now before the Supreme Court worth excluding innovation and competition, or the cost to administer a complex compensation system? The following section on the compensation debate delves more deeply into this tradeoff.

IV. COMPENSATION OF DEMAND RESPONSE IN ENERGY MARKETS

A. *The Order 745 Debate*

The core of the debate on compensation, now being considered by the Supreme Court, centers around two conflicting perspectives, that of the ISO, for whom a MW of generation or demand response is equivalent in terms of balancing the energy market, and that of the demand response providers,[27] who both receive the payment for DR deployed in the market and avoids the cost of the payment for energy never consumed. As discussed above, the gap in funding between what is paid to meet demand and what is collected by charging the marginal price to the actual demand creates a system cost which must be allocated to market participants.

Regarding the nature of the resource, numerous commenters addressed the physical or functional comparability of demand response and generation during the original consideration of this issue

[27] The 'demand response provider' as used here would include either the customer reducing its load, or the customer and its independent third-party load curtailment service provider, who together would receive the compound benefit.

by FERC, agreeing that an increment of generation is comparable to a decrement of load for purposes of balancing supply and demand in the day-ahead and real-time energy markets. Dr. Alfred E. Kahn stated:

> "[D]emand response (DR) is in all essential respects economically equivalent to supply response; and that economic efficiency requires, as the NOPR recognizes, that it should be rewarded with the same LMP that clears the market. Since DR is actually—and not merely metaphorically—equivalent to supply response, economic efficiency requires that it be regarded and rewarded, equivalently, as a resource proffered to system operators, and be treated equivalently to generation in competitive power markets."[28]

Dr. William Hogan, of Harvard, disagreed with FERC's intended approach in his comments, arguing that full LMP payment would be appropriate in some instances but not others.[29] Dr. Hogan distinguishes among three general types of demand response:

- Real-time Pricing Demand Response. Consumers are paying the applicable LMP for their marginal consumption.
- Explicit Contract Demand Response. Consumers purchase a fixed quantity of electricity but consume less than the purchased amount and sell back the difference.
- Imputed Demand Response. Consumers have an estimated consumption baseline and the difference between actual consumption and the baseline is the imputed demand response.[30]

According to Dr. Hogan, it is the last bucket of DR that causes the need for considering adjustment to full LMP. That is, Dr. Hogan agrees that the first two types of DR have a straightforward framework for using the applicable LMP.[31] However, he goes on to say that the Commission's policy is problematic because it is the third bucket that would be the most "ubiquitous" type of demand response affected by the policy.

The disagreement essentially stems from the perspective that the demand response product is equivalent to an unexercised call option on spot market energy, and the value of that option is well-established in finance theory as the value of the resource (Locational Marginal Price, or LMP) minus the "strike price," the retail tariff rate.[32] Payment of LMP without an offset for some portion of the retail rate fails to take into account the retail rate savings associated with demand response, and thereby over compensates the demand response provider. Demand response providers would "receive" both the cost savings from not consuming an increment of electricity at a particular price, plus an LMP payment for not consuming that same increment of

---

[28] Reply Comments of DR Supporters, Demand Response Compensation in Organized Markets, Docket No. RM10-17-000, August 30, 2010, Kahn Affidavit at p. 2 (Kahn Reply Affidavit).

[29] Hogan, William. "Demand Response Pricing in Organized Wholesale Markets." ISO/RTO Council Comments on Demand Response Compensation in Organized Wholesale Energy Markets 13 (2010): p. 1.

[30] *Id*. at p. 2.

[31] *Id*. These rate relationships assure the customer participating is not paid twice for the resource offered.

[32] Robert L. Borlick May 13, 2010 Comments at p. 4.

electricity.[33] If one then views LMP as a double-payment, As Hogan and EPSA does, paying LMP will in theory result in more demand response than is economically efficient.[34] This set of stakeholders hold that the demand response provider should therefore be compensated an amount less than LMP by the amount saved on the retail consumption of generated energy.

The primary error made by the supporters of what has come to be styled "LMP – G"[35] was to equate the opportunity cost of the customer with the lost value of electricity consumption, ignoring other costs and considerations. Quoting Dr. Kahn again, "the successful bidders for the opportunity to induce that consumer response are compensated for the costs of those efforts by the pool, whose (marginal) costs they save by assisting consumers to reduce their purchases."[36] Continuing, he notes:[37]

> "As to the remuneration of the successful bidders, promising reductions in purchases of power from the generators, their compensation from the pool itself is offset by the savings in (marginal) costs induced by their—successful—efforts to promote efficient demand response: that remuneration is emphatically not a net burden on generators (and not a subsidy), because it matches the savings in marginal generating costs that their successful efforts induce.
>
> "Observe that under effective competition, those middlemen are paid only once, not twice. To the extent they incur costs in inducing the demand response they promise, they will expect to retrieve them, profitably, from a combination of charges to consumers whose bills they have reduced and from the pool—the latter reflecting the savings in marginal generation costs they have effected."

Ignoring efficiency claims (which are more complex than advertised, due to existing distortions in retail and wholesale markets),[38] and focusing on the incentive effects on demand response providers, the issue becomes whether LMP or LMP–G will provide the optimal incentive to supply demand response. If indeed, the customer was merely reselling electricity in a purely

---

[33] Motion For Leave to Answer and Answer of the Electric Power Supply Association and White Paper by Professor William W. Hogan, Attachment A, Providing Incentives for Efficient Demand Response, Docket No. EL09-68-000, October 29, 2009, at pp. 15-19.

[34] See Comments of the Federal Trade Commission, Demand Response Compensation in Organized Markets, Docket No. RM10-17-000, May 13, 2010, at pp. 6-10; Comments of Potomac Economics, LTD., Demand Response Compensation in Organized Markets, Docket No. RM10-17-000, May 13, 2010, at pp. 6-8.

[35] Locational Marginal Price less the retail price to the load of Generation not consumed by the load offering the load reduction.

[36] Kahn Reply Affidavit, p. 10

[37] Kahn Reply Affidavit, p. 11

[38] There are numerous distortions in input markets to electricity production (subsidies, taxes, environmental regulations and mispricing of externalities), and widespread dislocations and distortions in virtually all economic aspects of relevant energy markets, as well as dislocations in the wholesale power markets, such as the existence of market power, imperfect information available to customers, barriers to entry and uneconomic resources dispatched to fulfill must-run requirements. For this reason, efficiency arguments should be limited to identification of gross distortions in pricing, and backed by empirical studies demonstrating that the impacts of these distortions are significant. Simplified economic models cannot provide definitive conclusions with regard to impacts on economic efficiency in a complex, "third best" world.

financial transaction with no risk as the option expired (i.e. the price was known with certainty), then LMP–G would be the optimal payment. However, if there are costs of providing demand response in addition to the lost opportunity cost, then LMP–G will be below the optimal payment, because the customer does not receive the full benefit of reducing its consumption.

Demand response is a physical, not a financial product that, similar to generation, incurs real costs and faces technical constraints. For aggregated retail customers, part of the payment is received by the demand response aggregator in return for its capital investment in equipment, operating costs and assumption of the risk of nonperformance. Many industrial customers face similar risks in curtailing operations as a generator does on start-up.[39] This suggests that the optimal price for demand response resources lies somewhere between LMP and LMP–G.

Furthermore requiring payment of LMP–G would in theory result in tracking retail consumption and rates for the multiple loads participating, as well as for their LSEs, and create undue confusion for retail customers and administrative difficulties for state commissions and ISOs and RTOs. In this case, if there are substantial benefits to demand response, and the administrative costs of attempting to implement LMP–G are similar in magnitude to G for most demand response resources, the best solution may be to charge LMP and simply uplift the cost on a load-proportionate basis.

As noted already, ERCOT pays the lower amount (LMP-G) to participating customers. So far there is very little participation by loads directly in SCED. Other factors limit the participation of loads in energy markets, as we noted, including requirements designed around the nature and characteristics of dispatch into an energy market, which restrict the participation of certain kinds of loads. Given the lack of participating loads, a higher payment based on LMP seems hardly likely to result in excessive demand response, but if this proved to be a concern, then restricting demand response resources from being accepted when their value is less than some threshold could work as an ad hoc corrective. In fact, a FERC adopted net benefit test attempts to do just this, limiting the times that loads can bid into the energy market to periods that can be expected to yield net benefits to all customers. In the real world, good policy may require administratively feasible balanced distortionary measures rather than an ideal policy that is theoretically economical, but administratively (or politically) infeasible. In economic policy, two wrongs may approximate a right.[40]

The existence of administrative costs and demand response related uplift (where total payments to generation and demand response exceed market revenues) provide an argument for limiting demand response payments to periods when the benefit to consumers was likely to exceed any deadweight costs transferred to consumers. In this case, demand response providers should be

[39] See Comments of Verso Paper Corp. in Support of the Notice of Proposed Rulemaking, Demand Response Compensation in Organized Markets, Docket No. RM10-17-000, May 13, 2010.

[40] A study of the PJM demand response program, pre-Order 745, which paid LMP above a price threshold of $75/MWh, showed a net social benefit. Whether this benefit would have been larger or smaller if LMP-G was paid was not determined. This suggests that whether LMP or LMP-G is the most socially efficient policy is an empirical question, one which unfortunately has yet to be thoroughly studied. Rahul Walawalkar, Seth Blumsack, Jay Apt and Stephen Fernands, An Economic Welfare Analysis of Demand Response in the PJM Electricity Market, *Energy Policy* 36 (2008): 3692–3702.

paid LMP only when the benefits of demand response compensation outweigh the costs to consumers as determined by some type of net benefits or cost-effectiveness test. Net benefits are most likely to be positive and greatest when the supply curve is steepest, which typically occurs in highest-cost, peak hours.

While the impact of demand response resources on the energy market is the same as generation on the margin, this does not mean demand response is equivalent to generation. Traditional generators provide system support features that demand response cannot, such as governor response and reactive power voltage support, which are necessary for reliable operation of the electric system. This would suggest that there may be diminishing social returns to economic demand response at some level of market penetration. Limiting the hours in which demand response resources are paid LMP could also help establish better baselines for measuring whether a demand response provider has, in fact, responded.[41]

Given the desirability of a net benefits test, the next question for FERC was what kind of test? Some commenters advocated a net benefits trigger based on a particular price threshold, which is easy to implement. However, using a static threshold based on historical data misses the changes that occur within electricity markets across seasons and years. A price threshold could be based on a preset heat rate of marginal generation and fuel price, like that currently used in New England's Day-Ahead Load Response Program and ERCOT's bid caps. Other commenters suggested a dynamic cost/benefit analysis built into the dispatch algorithm. However, a dynamic net benefits test done on an hourly basis would become very complicated to implement.[42] The FERC settled on a monthly analysis to identify a price threshold where customer net benefits would occur.[43]

The Technical Advisory Committee (TAC) of ERCOT voted to endorse LMP-G rather than LMP as the mechanism to enable direct participation in the real-time market in 2011. As presented to TAC, LMP-G establishes the principle that a customer should not get the benefit of the curtailment twice--i.e., LMP plus avoided cost of energy. Neither the ERCOT Board, nor the PUCT has voted on the issue of Full LMP versus LMP-G as of this writing. However, ERCOT staff and stakeholders are still struggling to develop procedures to implement LMP-G for third-party providers that are feasible and don't violate Texas law or the Commission's rules. It turns out that implementing LMP-G presents complex issues that will result in significant administrative costs and/or the application of ad hoc "fixes" that will result in a price that still is unlikely to meet the theoretical ideal for economic efficiency.[44]

[41] See Comments of the State of New York Department of Public Service, The New England Conference of Public Service Commissioners, Maryland Public Service Commission, New York Public Service Commission, in Demand Response Compensation in Organized Markets, Docket No. RM10-17-000, May 13, 2010.
[42] Order 745 at PP 68-76.

[43] Order 745 at P 79.
[44] "LMP-G in ERCOT" White Paper of the Loads in SCED2 Subgroup of the Demand Response Working Group of ERCOT, July, 2015

B. *Demand Response, Capacity Markets and Energy Only Markets*

Another concern of some stakeholders is that economic demand response will reduce peak energy prices, increasing the "missing money" problem[45] that is at the core of the resource adequacy issue in "energy-only" markets. In electricity markets with corresponding capacity markets, generation owners may raise capacity price offers to remain financially viable at lower energy prices. Demand response should flatten the load profile and decrease the forecast of load growth projections, which would reduce capacity clearing prices. It becomes an empirical question whether and to what extent energy price reductions will result in capacity price increases. For purposes of long-run reliability, as long as compensation is sufficient to induce new investment and reflects market value, demand response in the bid stack will only push out high cost generators.

In an energy-only market such as ERCOT, the question is slightly different. Since there is no capacity market, the issue becomes the impact on long-run energy prices. These savings are primarily a transfer of infra-marginal rents from generators to consumers. The impact on long-run energy price will reflect the extent to which new capacity relies on price spikes as a source of infra-marginal rents to justify investment in new facilities. If price spikes are highly variable, then they will be discounted by investors (because of the increased risk due to the uncertain timing of revenues streams) and much of the benefit of these spikes accrue to existing generators as economic rents. To the extent that demand response moderates price spikes and makes energy revenues more predictable, investors will accept a lower rate of return given the same expected energy revenue. Thus while long-term energy prices will rise to some extent to compensate for lost revenue due to demand response resources, this price increase will be less than that needed to completely recover the lost revenue. So, an increase in demand response resources could also reduce the optimal level of generation capacity, further reducing the required long-term price increase.

C. *Other Considerations*

A final consideration is that the lack of responsive demand at the margin contributes to price volatility in electricity markets and establishes conditions conducive to the exercise of market power. While the price impact of demand response may have been overstated (because of the missing money problem, reducing revenues to generators will result in some rebound in capacity market prices or long-term energy prices), the importance of demand response resources as an automatic check on the exercise of market power has generally been understated. Since any exercise of market power (raising price above theoretically efficient levels) creates both economic losses and unjustly transfers income from consumers to generators, demand response can act as a counter-balance. While the FERC and the ISOs have improved their market

[45] Because competition drives resources to offer at their marginal operating cost, economists and regulators worry that the market may not provide sufficient volatility to insure a return on capital investment and encourage the commitment of new resources as they are needed. This added incentive has come to be labeled the "missing money."

monitoring and mitigation efforts, these are designed to ensure that markets are "workably competitive" that is, market power is restricted to a tolerable level (which is not a criticism, excessive mitigation creates its own sets of problems and inefficiencies). Economic demand response does not interfere with mitigation of market power, rather, like forward contracting, it makes it more difficult and less lucrative to attempt to exercise market power independently of market mitigation efforts.

V. CONCLUSIONS

The controversy over demand response resource compensation in energy markets reflects disagreements over economic theory, the problem of identifying feasible economic policies, and self-interest.[46] The underlying problem is that most customers face flat rates, or flat rates with limited peak pricing features, and thus do not even have the opportunity to properly respond to price signals. Even here in ERCOT where the opportunity exists, it appears insufficient alone to capture the potential efficiency available. Inclusion of loads willing to participate in the price formation process represented by the operation of the wholesale market is a venue to address this inefficiency. The emergence of a new class of services, and service providers, to provide the marketing, technology, knowledge and assistance customers need to be induced to participate, seems to be the vehicle.

The optimal compensation for economic demand response resources, and what limitations should be placed on demand response bidding into energy markets, is an open question. The optimal compensation probably lies between LMP and LMP-G, while some sort of threshold trigger might be appropriate if there is a danger of overcompensation with using LMP, but not with LMP-G. Conversely, if the administrative cost and complexity of using LMP-G is substantial, it may be more "efficient" to use LMP. There has been too much effort expended to determine the "perfect" formula, and too little effort expended gathering empirical data on how different demand response resources respond to different incentives. Given the limited quantity of demand response resources actively participating in energy markets, the optimal short-term strategy may be to err on the side of potentially excessive compensation, gather data on demand response resources and their impact on electricity markets, then revisit the issue when there is sufficient experience.

History shows that regulators often have to muddle through to find the best policies for governing electricity markets.[47] In the case of demand response, the stakes are far smaller than many other electricity market policies, from resource adequacy mechanisms to market mitigation. Therefore, a policy of encouraging demand response in electricity markets to a level of penetration that would allow determination of the effectiveness and relative efficiency of this

[46] Demand response resources obviously want to maximize their compensation, aggregators their net revenue, while generators want to block competition from resources that could reduce energy prices. See Steven Salop and David Scheffman, Raising Rivals' Costs, *American Economic Review* 73 (1983): 267-71. It is important to keep this in mind when reviewing the arguments of their experts.

[47] Steve Isser, *Electricity Restructuring in the United States*, at 459-63.

resource may be preferable to a conservative approach that denies the market this opportunity. At worst, this policy could encourage investment in infrastructure and institutional experience that might provide additional resources for ancillary service markets and emergency demand response if experience counsels less generous compensation in energy markets.

Appendix A: Incremental Demand Response Analysis: ERCOT Case Study

Summary of Study

Demand Response (hereafter, DR) is a term used today to describe a reduction or shift in electric demand in response to market signals or incentives. It can involve changes to industrial or commercial processes, or adjustments to energy using equipment, or, if permitted, the use of on-site generation or storage. DR can be "economic," dispatched in response to price signals, or responsive to emergency or reliability conditions, as determined by a utility, competitive retailer, or system operator. It can be achieved individually or with assistance, sometimes as part of a managed aggregation.

ERCOT, the Electric Reliability Council of Texas, is the independent system operator (ISO) internal to, and serving the majority of, Texas.[48] As such, it manages the real-time flow of electricity across the transmission system to maintain the safe and reliable delivery of electricity to Texas consumers. In its capacity as the ISO, ERCOT also operates the wholesale market which features locational marginal pricing (LMP) for generation at more than 8,000 nodes, a day-ahead energy and ancillary services co-optimized market, day-ahead and hourly reliability-unit commitment, and congestion revenue rights.[49] This competitive market continuously matches buyers and sellers to determine the wholesale price of electricity in 15-minute intervals. The price tends to climb as demand for electricity rises, often because older, less efficient generation plants must be called upon for infrequent high demand periods. Strategic reductions of consumer demand, acting as resources for the system, especially during periods of high demand or emergencies, can eliminate the need for these marginal resources and thereby reduce the marginal cost of power for any given increment of time. DR may also reduce environmental impacts of the power system.

The purpose of this study is to estimate the financial impact that stimulation of additional demand response (DR) would have on Texas consumers. To do that, we have used the available historical data to estimate the likely impact of load reductions within the ERCOT footprint on several critical days between 2011 and 2013. In the case of years 2012 and 2013, available data allowed us to estimate potential savings from this incremental economic DR, by determining what the system average LMP would have been if there had been various levels of additional supply from DR resources. In this context, "economic DR" is DR by customers that would be willing to voluntarily respond by removing load from the ERCOT system for an acceptable price. To estimate the potential savings, we examined the actual price offers of generation on several specific days during years 2012 and 2013 that exhibited intervals of sustained marginal prices above assumed 'strike prices,' or offer prices for DR. That is to say we identified the specific market intervals during which average locational prices were high enough that it would be reasonable to assume that given the opportunity, some loads would readily opt to curtail their energy consumption in order to earn the market price for wholesale electricity. We then examined the available generation offer prices (supply curve data) for all resources offered into

[48] ERCOT is a private corporation formed by stakeholders under the authorization of the legislature and under the oversight of the Public Utility Commission. ERCOT both oversees the planning for transmission interconnection of the region and administers wholesale energy and ancillary services markets.

[49] LMP is the offer-based marginal cost of serving the next increment of Load at a given network node.

the ERCOT wholesale market during those intervals, and added our assumed quantities of additional DR at the assumed strike prices to the supply curves.[50] Then, holding average system demand for the modeled hours constant, we estimated what the system average price would have been by observing the price at which the offered quantities from the modified supply curves intersected system demand. In other words, we backed out of the supply curve an amount of generation equivalent to the amount of DR offered at a lower price.

We then calculated the total savings associated with the new, now lower marginal price, at the original level of system demand.[51] The lower marginal price is not only applicable to the marginal generation unit(s) still needed to serve load, however: all units dispatched by ERCOT for the increment of time considered would have also been paid the lower marginal prices in the applicable intervals–-leveraging potentially substantial overall customer savings.[52] For our evaluation we assumed that all energy was traded in the real-time (spot) market on the days in question, but we recognize that the actual potential savings would be reduced in proportion to the energy traded in forward contracts at prices below the marginal price set by the market for that interval.

In the case of 2011, at the time of this study ERCOT had not published equivalent 2011 historic data as it has for 2012 and 2013.[53] This level of data is a requirement to model the impact of additional economic DR in ERCOT's energy market auctions. Fortunately, however, we did have sufficient 2011 data to model the impact of additional emergency DR in severe system conditions as existed in 2011 but did not occur in 2012 or 2013.[54] We assumed that the additional emergency DR would have reduced the level of involuntary load shedding during periods of supply shortfall. For this part of the study, we utilized Value of Lost Load (VOLL) estimates as included in a London Economics study commissioned by ERCOT.[55] VOLL is the value that theoretically reflects a customer's "willingness to pay" for reliable electricity service. It is generally measured in dollars per unit of power (e.g., per megawatt hour, "MWh"). As noted in the London Economics study for ERCOT, accurately estimating VOLL for a given region and a specific type of outage is a challenging undertaking–as VOLL depends on multiple

[50] Supply curves are upward sloping, representing the positive relationship between willingness to supply and price. That is, at higher prices, suppliers will be willing to increase their quantity supplied, or additional higher-priced resources are able to offer into the market.

[51] Actual system demand would have been reduced by the amount of incremental DR that would have cleared in the modeled hours. We applied the new, lower, marginal price to the original demand in order to account for the market's payment for the cleared DR.

[52] Units included in the dispatch to serve load, but not at the "margin" setting the marginal price are referred to as "infra-marginal."

[53] The issues were communicated by ERCOT in market notices (W-B120710), which included the 48-hour Aggregate Supply Curve for Non-Wind Resources report and the 48-hour Aggregate Supply Curve for Wind Resources.

[54] For example, using the February 2, 2011 firm load shedding information and timeline, we assumed that additional DR Emergency Response Service), in 500 MW increments, would have reduced the amount of firm load shedding by the same amount. We then multiplied the MWh of additional deployed DR by an assumed Value of Lost Load (VOLL) to arrive at an estimated savings to consumers.

[55] There is no single agreed upon VOLL figure for ERCOT, we assumed that VOLL fell within the range provided in, ***Briefing paper prepared for the Electric Reliability Council of Texas, Inc. by London Economics International LLC***, June 27, 2013, and that ERCOT would not set a price cap higher than their estimation of VOLL and, therefore, the current price cap is a reasonable (and probably a conservative) proxy for VOLL.

factors such as the type of customer affected, regional economic conditions and demographics, time and duration of outage, in addition to other specific traits of a given outage.

2012 and 2013 Findings

As noted above, we set out to analyze the impact of economic DR on several days during relatively mild summers (2012 and 2013). We found that including 1500 megawatts of economic DR—approximately 2.5% of total demand—would bring potential savings of as much as $200 million on only the five days we modeled.

Available data allowed us to estimate potential savings from incremental "economic DR," by identifying the lower marginal price due to incremental economic DR availability.[56] To estimate the potential savings, we selected several days during years 2012 and 2013 that exhibited intervals of sustained marginal prices above an assumed "strike ("offer") price" for DR.[57] That is to say, we identified several market intervals during which average locational prices were high enough that it would be reasonable to assume that given the opportunity, some loads would readily opt to curtail their energy consumption in order to earn the market price for wholesale electricity. We then examined the available supply curve data for all resources offered into the ERCOT wholesale market during those intervals, and added to the supply curves our assumed quantities of additional DR at the assumed strike prices. Then, holding average system demand for the modeled hours constant, we estimated what the system average price would have been by observing the strike price at which the offered quantities from the modified supply curves intersected system demand.

Savings were then calculated by utilizing the new, lower marginal price inclusive of the incremental economic DR at the original level of system demand. Of course, not only is the lower marginal price applicable to the marginal generation unit(s) still needed to serve load, all infra-marginal units dispatched by ERCOT for the increment of time considered would have also be paid the lower marginal prices in the applicable intervals – leveraging potentially substantial overall customer savings.[58]

Although this analysis focuses on only 5 days during 2012 and 2013, the estimated savings associated with incremental, economic DR are potentially substantial. Utilizing the methodology described above, our analysis estimated that incremental, economic DR could have resulted in wholesale market savings of over $200 million during the examined intervals of high prices in ERCOT during 2012 and 2013. As noted, this total potential savings assumes all energy was traded in the spot market during the intervals examined, and so the total would be reduced in proportion to the quantities traded in forward curves below the marginal price that would have been struck in the absence of the DR contribution. We also recognize that the introduction of 1500 MW of DR could cause longer-term price impacts that could moderate the spot market value of the incremental resource. Nevertheless, the examination of the actual market bids

[56] We worked with publicly available ERCOT data from reports published pursuant to ERCOT's Protocols.

[57] We utilized three tranches of DR offers, 500 MW at $300/MWh; an additional 500 MW at $500/MWh; and an additional 500 MW at $1,000/MWh.

[58] Infra-marginal refers to the units inside of, as opposed to at the margin.

provides a realistic indicator of the level of value of an incremental amount of demand response.

The following table presents the estimated Real Time Market (Security Constrained Economic Dispatch) savings for the spot market intervals noted.[59]

SUMMARY OF SAVINGS IN 2012 AND 2013

	3/31/2012	4/26/2012	6/26/2012	9/3/2013	10/1/2013	Total
Hours of DR Deployment	2	1	3	1	2	9
MWh of DR Deployment	431	597	2986	1095	1362	6470
Savings from DR Deployment	$52,151,210	$2,907,447	$83,421,969	$37,516,800	$28,077,316	$204,074,742

Cost implications of Economic DR

As noted above, our analysis included the assumption that the market would pay for the participating DR at the full price of energy for the interval in which the contributions were economically beneficial. In only two hours were some or all of the megawatts with a $1000/MWh strike price accepted. As discussed at length in "The Debate about Demand Response," to which this is appended, the issue of the proper compensation of DR has risen to the consideration of the Supreme Court this year. The Federal Energy Regulatory Commission directed wholesale markets over which it has authority to pay the full locational marginal price for DR. Opponents argue that this is an over compensation, because the customer also received the benefit of the avoided consumption charge. To offer an energy reduction into the market a customer must first purchase that energy, otherwise it is simply a retail decision not to consume. So, it follows from this logic that the appropriate compensation by the wholesale market would be the LMP minus an amount equivalent to the retail price of electricity to that contributing load. This lower payment has come to be labeled "LMP-G," with G representing the generation cost component of the payment.

The theoretical, sometimes ideological, debate about this is so intense that we also wanted to better understand what would be the actual impact of a policy of LMP versus LMP-G if implemented. Using the data we obtained for the actual bid stacks during the 9 hours of DR evaluated above, we found that paying the full LMP to achieve as much as $200 million in reduced energy costs would cost about $4.3 million at full LMP, and assuming the load paid 7.5 cents per kWh ($75/MWh) for its electric rate, about $3.8 million at LMP-G, for a difference of about $500,000. Even at a retail rate of 30 cents per kWh ($300/MWh) for its marginal consumption, the difference to the market would be just under $2 million, or one percent of the total potential savings. So, even considering our savings estimate would be moderated by out of market trades and future price adjustments by remaining generation, this is a very small cost in comparison to the potential savings.

[59] March 31, 2012, Hour Beginning (HB)16 and HB17; April 26, 2012, HB16; June 26, 2012, HB14, HB15, HB16; September 3, 2013, HB16; and October 1, 2013, HB15, HB16.

2011 Findings

2011 was a record setting year in ERCOT, with a new peak demand record of 68,379 MW on August 3, 2011. In fact, the 2010 peak demand of 65,776 MW was broken on three consecutive days: Aug. 1, 2011 peak demand = 66,867 MW, Aug. 2, 2011 peak demand = 67,929 MW, and Aug. 3, 2011 peak demand = 68,379 MW. ERCOT also experienced a new weekend record on Sunday, Aug. 28 of 65,159 MW: an increase of 5% over the previous record. In addition, ERCOT set a new winter peak record of 57,282 MW on February 10, 2011. We are confident that our estimate of savings in the mild weather years of 2012 and 2013 would be dwarfed by the savings potential in 2011.

In examining the 2011 data that was available, we found a February 2nd event of firm load shedding (FLS). During the early morning hours on this day, ERCOT experienced extreme cold weather, record electricity demand levels, and the loss of numerous electric generating facilities.[60] This firm load shedding event lasted 7 hours and 25 minutes, during which several "blocks" of load were shed.[61] To estimate the value of incremental "emergency" DR that could have been deployed during this event, we assumed that the cost of an additional 500, 1000, and 1500 MW was the same average $/MW-day value as the capacity that ERCOT actually purchased for delivery year 2011.[62] We then estimated the savings that might have accrued by applying three conservative tiers of Value Of Lost Load (VOLL) to the assumed additional DR that we modeled to offset firm load shed.

For our low-end estimate of VOLL, we chose the 2015 offer cap value for ERCOT: $9,000/MWh. We believe this is a conservative low-end estimate, based on the London Economics Study and our assumption that ERCOT and the PUCT would not implement a price cap that is higher than their estimate for VOLL. Our analysis showed that regardless of our VOLL assumptions, there is a positive net benefit for 2011, and in several scenarios there is a positive net benefit over multiple years. Table 1 below shows the additional DR cost as a percentage of savings to consumers–the bang for the buck, so to speak. In all scenarios, and at all deployment levels we used, consumers would receive benefit greater than the cost of purchasing the emergency DR.

Table 1.

Additional Emergency DR Cost as a % of Savings to Consumers: 3 VOLL Scenarios			
	Additional DR (MW)		
Scenario	500 MW	1000 MW	1500 MW
1: VOLL = $9,000	77%	79%	82%
2: VOLL = $17,967	39%	47%	41%
3: VOLL = $26,953	26%	26%	28%

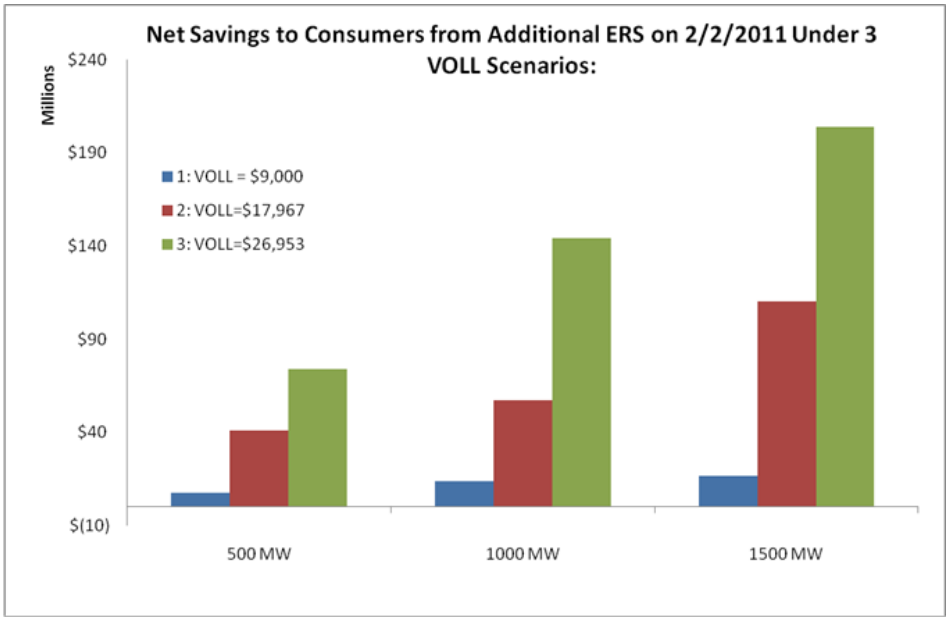
[60] IMM Report to the PUCT, April 21, 2011, p. 1

[61] Energy Emergency Alert (EEA) Level 3 was declared by ERCOT at 5:43 a.m. All firm load was finally restored shortly after 1:00 p.m.

[62] In our analysis, we ignored the current, arbitrary $50MM per year spend cap for ERS because the gross spending cap clearly prohibits substantial additional net savings.

Table 2 below summarizes the net savings under the 3 different VOLL scenarios to consumers from additional emergency DR utilization during the firm load shedding event on February 2, 2011.

Table 2.



Not surprisingly, the savings are even more striking at higher assumed VOLL. As a second tier of analysis, we choose to utilize as our mid-level assumption for VOLL the high-end of the residential class as discussed in the London Economics report of VOLL in ERCOT. This level showed savings in a range between $40.8 million for 500 MW of DR, to $110 million for 1500 MW of deployed DR.[63] Finally, we chose to model a high-end for VOLL representative of the value for commercial and industrial customers. For this we used half of the high end VOLL for Industrial customers as discussed in the same London Economics report. As is evident in the table above, 1500 MW of deployed DR at a VOLL of $26,953 would have resulted in avoided costs associated with firm load shedding of $203.8 million.[64]

Thus, the range of estimated net savings for this single firm load shedding event is $7.6 million – with VOLL equal to $9,000 and deployed DR equal to 500 MW – to $203.8 million – with VOLL equal to $26,953 and deployed DR equal to 1500 MW.

Net Savings from Additional Emergency DR: 3 Scenarios			
Scenario	**500 MW Additional DR**	**1000 MW additional DR**	**1500 MW Additional DR**
1: VOLL = $9,000/MWh	$7,585,925	$13,746,850	$16,532,775
2: VOLL = $17,967/MWh	$40,838,550	$57,057,325	$110,088,475
3: VOLL = $26,953/MWh	$74,161,633	$144,055,708	$203,842,408

[63] "Savings" in the context being used here accrues due to the avoidance of firm load shedding due to the addition of incremental, voluntary DR deployment.

[64] This VOLL is roughly ½ for the customer class in the London Economics study performed for ERCOT.

Conclusion

Theoretical arguments abound regarding the effect of active demand side participation in wholesale electric markets, but our analysis demonstrates the real value that such active participation could bring.[65] In very few days, and with very conservative assumptions, our analysis shows potential savings that must be considered significant. A modest number of consumers strategically reducing their demand for electricity in response to price signals on only five total days in 2012 and 2013 could have reduced power costs for all consumers in the state more than $200 million. The technology to facilitate this active participation by demand side resources is becoming increasingly prevalent, and will allow virtually all loads to actively participate in ERCOT wholesale markets.

Although data were not available to estimate the total savings from incremental, economic DR in 2011, as was the case in 2012 and 2013, we were able to estimate the savings associated with incremental emergency DR in 2011. ERCOT experienced extremely cold winter weather, and extremely hot summer weather in 2011. No doubt, during this extreme year several intervals of high demand would have resulted in extremely high market prices; and based upon are findings in 2012 and 2013, incremental, economic DR would have been poised to participate and generate substantial savings.

Specifically, our analysis showed that on the single firm load shedding day, February 2, 2011, depending on the level of DR participation, and the VOLL assumed, the estimated net savings that could have been realized ranged between $7.6 million and $203.8 million: by any account a significant cost associated with a firm load shedding event that might have been mitigated by additional incremental DR.

All told, depending on which VOLL we use for 2011, on 6 days during 2011, 2012, and 2013, consumers could potentially have saved between $226MM and $422MM over the days we modeled in those years.[66] Taking the midpoint of conservative estimates of savings and VOLL, $110 million in costs associated with firm load shedding could have been avoided by active DR participation. Added to the estimated savings associated with incremental, economic DR in 2012 and 2013, this adds up to a potential savings for ERCOT consumers of up to over $300 million – in 6 scant days!

[65] See, for example, Schweppe, Fred C., Michael C. Caramanis, Richard D. Tabors, and Roger E. Bohn. Spot Pricing of Electricity, Boston: Kluwer Academic, 1988. Print.

[66] The six days are equal to five days of incremental, economic, real time DR in 2012 and 2013, and one day of emergency DR in 2011.

This page is intentionally left blank

The South-central Partnership for Energy Efficiency as a Resource

www.EEPartnership.org