Demand Management in Restructured Wholesale Electricity Markets

Hung-po Chao
ISO New England

May 20, 2010

Table of Contents

Demand Management in Restructured Wholesale Electricity Markets

Hung-po Chao[1]
May 20, 2010

SUMMARY

Electric restructuring has created a hybrid market structure in which wholesale market prices are set at competitive levels, but in large part, retail rates are still fixed at the average cost of service. A critical lesson from the California electricity crisis of 2000 and 2001 shared among many experts is that competitive electricity markets are not sustainable with a hybrid market structure, if the retail prices that electricity consumers pay are completely disconnected from wholesale market prices. Following the Congressional mandate in the Energy Policy Act of 2005, federal and state regulators have promoted demand response programs in various regions. Recently, FERC issued a proposed rule for demand-response compensation for the purpose of promoting price-responsive demand in organized wholesale electricity markets.

This paper examines alternative approaches for demand response in the wholesale electricity market focusing on the interactions between competitive wholesale prices and regulated retail rates. The traditional approach, which is inherited from the practice before restructuring in the vertically integrated utility depending on the use of administratively determined customer baselines, has been commonly adopted by existing demand response programs. Unfortunately, this approach may produce unintended effects, including baseline manipulation, inefficient price formation and incentives for generation relocation and load shifting behind the meters, mainly because consumers do not own the energy for which they are paid pursuant to demand response programs. Fundamentally, absent energy entitlement or property right for the baseline, paying the full wholesale energy price for demand reduction can compromise electricity market efficiency and make the cure worse than the disease.

Demand subscription is one possible contract-based retail rate design remedy for the misaligned incentives associated with demand response programs that rely on administratively determined customer baselines. Demand subscription establishes the contractual rights and obligations for the customer baseline that enables an efficient demand response program. As a risk sharing mechanism, demand subscription provides appropriate incentives by only allowing electricity consumers to sell energy that they have made a firm commitment to purchase. In addition, making available risk management options via a menu

[1] Director, Market Strategy and Analysis, ISO New England Inc., One Sullivan Road, Holyoke, MA 01040. I greatly appreciate the comments from many colleagues and the participants at the HEPG meeting held at Harvard Kennedy School on May 20-21, 2010. The views expressed herein are those of the author and do not necessarily represent the positions of the ISO or other organizations.

of differentiated services, demand subscription encourages electricity consumers to opt into real-time pricing tariffs, promoting price-responsive demand in wholesale and retail electricity markets.

1. Introduction

Electric industry restructuring over the past three decades has created a hybrid market structure. On the one hand, competitive wholesale markets produce market clearing prices that generally reflect the time-varying marginal cost of generation. On the other hand, retail rates charged to the vast majority of consumers retain the fixed rate structure inherited from the vertically integrated utility that existed before restructuring. These rates do not dynamically reflect time-varying marginal generation costs, but tend to reflect the average cost-of-service over a period ranging from a few months to perhaps several years. Further, these fixed retail rates generally provide customers an implicit call option by giving them the right to consume unlimited amounts of electricity at a fixed price.[2]

Price-responsive demand generally refers to the idea of changes in end users electric consumption in response to changes in the price of electricity. Price-responsive demand can be enabled through many different mechanisms. Dynamic pricing, i.e., retail tariffs that either directly reflect wholesale real-time prices or at least reflect critical peak conditions, is generally considered an efficient way to enable price-responsive demand. However, because of significant regulatory and technological barriers to dynamic pricing, the application of dynamic pricing has been comparatively limited.

In contrast, demand response refers to the reduction in end user's electricity consumption in response to the price of electricity or other market incentives.[3] In recent years, demand response programs have been developed extensively in wholesale electricity markets. Overlaying consumers' retail tariffs, demand response is essentially a derivative product that

[2] The unlimited quantity option with price hedge was introduced when the industry supply was characterized by economies of scale with declining costs. However, from a consumer's perspective, such an option could be fairly expensive in the presence of diminishing scale economies with increasing supply costs.

[3] In general, the terms "demand response" and "price-responsive demand" have been used interchangeably, though there are important distinctions in practice. Price-responsive demand places a greater emphasis on using energy market price signals to improve the economic efficiency of energy consumption. Demand-response programs in wholesale markets refer predominantly to demand reduction or curtailable service programs designed to improve system reliability. Demand-response programs focused on reliability attempt to provide capacity or reserves to the electric system by paying customers to take their electrical load off the bulk power system when the system is deficient in capacity or operating reserves. Demand-response programs in retail markets are predominantly time-based pricing programs, including real-time pricing and other dynamic retail pricing programs. In the following, demand response primarily refers to the "supply-side" approach of paying for demand reductions in the wholesale markets.

provides incentive payments for reductions in consumption relative to the counterfactual consumption level absent the demand response payments. One of the common characteristics of the existing demand response programs administered by many ISO/RTOs is their dependence on the use of an administratively-determined customer baseline. The customer baseline is an estimate of an end user's consumption level in the absence of incentive payments to reduce his consumption. The use of an administratively-determined customer baseline introduces several unintended effects: baseline manipulation, inefficient price formation, and generation relocation and load shifting behind the meters. If these problems are not properly addressed, demand response programs could be counterproductive and deter efficient development of price-responsive demand, likely making the cure worse than the disease.

The problems introduced by the use of an administratively determined customer baseline are rooted in the traditional vertically-integrated utility system. Given the structure of most demand response programs, the use of an administrative baseline provides consumers the opportunity to sell electricity that they have not made firm commitments to buy and do so at wholesale prices that fail to exceed consumers' marginal value of consumption. As we will demonstrate below, this could distort price formation in the wholesale market and create perverse incentives for gaming, resulting in illusory demand response.

Demand subscription, a contract-based retail rate design approach, addresses the customer baseline problem via a two-settlement transactions system that allows the retail customer to contract *ex ante* for a specific amount of energy service and take ownership of the baseline. In general, demand subscription transforms an unlimited quantity option into a spectrum of priority service call options that may be triggered when the real-time price exceeds pre-specified levels. As a result, demand subscription eliminates the need for using administratively-determined customer baselines. Instead, the subscribed level of demand naturally defines a two-sided customer baseline, working like a two-part real-time pricing program, and provides a superior risk sharing mechanism for consumer to align the economic benefits by shifting consumption from peak hours to off-peak hours. As a result, demand subscription encourages electricity consumers to opt into real-time pricing tariffs, promoting price-responsive demand in wholesale and retail electricity.

The purpose of this paper is to present an exposition of public utility economics that bears on the demand management issues in a hybrid market structure. Conceptually, this paper draws on the economic framework of peak-load pricing theory that incorporates the interactions between competitive wholesale prices and regulated uniform retail rate within a hybrid market structure.[4] This paper begins with some basic questions: What is the nature or product

[4] See Chao, H. (1983) for a demonstration of the relationship between optimal retail rates and competitive wholesale prices under demand and supply uncertainty. For a survey of the peak-load pricing theory, see Crew, Fernando and Kleindorfer (1995).

definition of demand response? What are the benefits of demand response? What is the rationale for demand response in the wholesale electricity markets? To illuminate the issue, we evaluate alternative approaches to integrate demand response in the wholesale electricity market based on existing proposals. Simple examples are used to elucidate the main insights.

The remaining sections are as follows. In Section 2, we review the rationale for demand response in wholesale electricity markets. In Section 3, we examine the alternative approaches to compensating demand-response resources in restructured wholesale electricity markets. In Section 5, we estimate the economic benefits of alternative approaches. In Section 6, we conclude with a summary of the key insights.

2. The Rationale for Demand Response in Wholesale Electricity Markets

Historical Background

The Public Utility Regulatory Policies Act (PURPA) of 1978 opened an opportunity to introduce competition into electricity generation that led to restructuring of the electric industry. This process was facilitated by the removal of the restrictions on the use of gas in electricity generation under federal law and the technological breakthrough of combined-cycle gas turbines. As a consequence, electricity generation was no longer characterized by economies of scale, and the assumption that electricity generation was a natural monopoly was no longer valid, though the transmission and distribution systems still exhibit economies of scale and remain natural monopolies.

The Energy Policy Act of 1992 and subsequently the FERC Order No. 888 opened access by non-utilities and independent power producers to the transmission networks.   On the state level, the electricity restructuring process was initiated on April 20, 1994, when the California Public Utility Commission (CPUC) issued its *Blue Book*. As envisioned in the *Blue Book*, electricity restructuring, in essence, entails an evolution of the traditional regulatory compact. The traditional regulatory compact is composed of several key elements. First, it grants the utility monopoly franchise rights. Second, it grants the utility an opportunity to recover reasonably incurred costs and earn a fair return on its investment. Third, in return, the utility is subject to regulation to provide safe and reliable service, priced fairly and reasonably, to all consumers within its monopoly franchise. Vertical integration of public utilities resulted from the strong public interest in the safe and reliable provision of an essential service and the dominant technological and economic characteristics of electricity industry, including natural monopoly, economies of scale, economies of scope and high transaction costs with incomplete metering data. Restructuring reflects the underlying changes in market structure and progress in technological development. Restructuring shifts the rate design approach from cost-of-service rates to market-based rates for improved performance incentives.[5] By expanding competition in power generation, restructuring was expected to create incentives for cost cutting and encourage investment in efficient new generating capacity. In general, the competitive market is intended to serve as an efficient means to achieve the given public policy objectives.

A critical lesson from the California electricity crisis of 2000 and 2001 shared among many experts is that competitive electricity markets are not sustainable with a hybrid market structure, in which consumers are completely disconnected from wholesale market prices.[6] The disconnection became apparent during the crisis, as demand for electricity appeared price insensitive or inelastic.[7] Dynamic retail pricing is generally viewed as an efficient way of encouraging price-responsive demand, but there are well-recognized strategic barriers to dynamic pricing and price-responsive demand. [8] Since 2001, demand-response programs in wholesale electricity markets have been created in part to overcome such barriers.

Price-responsive demand can improve economic efficiency, reduce reliability costs, and produce environmental benefits. Moreover, active demand participation in wholesale energy markets mitigates the potential exercise of market power. However, the electricity restructuring process in California failed to develop price-responsive demand that would respond to wholesale price spikes by reducing consumption. During the California electricity crisis in 2000 and 2001, producers exploited the disconnected market structure with inelastic demand by raising prices. Few restraints were in place to curb such exploitative behavior, and the utilities selling power to retail customers subject to retail price caps but without long-term contracts were caught in-between with few options as wholesale power prices skyrocketed. In retrospect, it was obvious that if the demand for electricity had been responsive to changes in the wholesale price, the attempt to raise prices would have caused a reduction in sales, thus reducing the profitability of such an action and discouraging the exploitative market behavior.

The Energy Policy Act of 2005 provides a Congressional mandate for demand response in organized wholesale electricity markets and directs FERC to "work with States, stakeholders and experts to identify and address barriers to the adoption of demand-response programs" and "encourage States to coordinate, on a regional basis, State energy policies to provide reliable and affordable demand-response services to the public." While there are many

[5] See Chao, H., Oren, S. and Wilson, R. (2008).

[6] See Joskow (2001), Hogan (2001), Sweeney (2002), Faruqui, et. al. (2001)

[7] See Joskow (2001) and Hirst and Kirby (2001).

[8] See FERC (2006).

difficult barriers, the benefits of price-responsive demand are many. (FERC 2006, 2009)[9] Federal and state regulators have endorsed demand-response programs that pay consumers to reduce their demand during high price periods.[10] On March 18, 2010, FERC issued the proposed rule on demand-response compensation in organized wholesale electricity markets (NOPR). The proposed rule requires the ISOs and RTOs to pay demand-response providers, *in all hours,* the wholesale energy market price for demand reductions.[11]

Lowering the Barriers to Price-Responsive Demand

Demand response refers to a customer's ability to alter its electricity demand by reducing or shifting consumption in response to the price of electricity or other market incentives. As a resource in the wholesale market, demand response can be thought of as a composite *option*, or essentially a derivative product, with four relevant components: 1) A customer baseline (CBL), which is established when the unlimited quantity option in the basic electric service is exercised, e.g., when the customer buys the baseline 2) a wholesale price threshold (Threshold), 3) an offer price (Offer) in the wholesale energy market, and 4) a compensation rule for the demand reduction, which is the difference between the actual consumption (Q) and the baseline, when certain conditions are satisfied (e.g. Q < CBL, LMP > Threshold, LMP > Offer). Before the option is exercised, demand response resource can deter the exercise of market power by providing price elasticity in response to high market prices. In this sense, demand response resource is a capacity product (*ex ante*) that provides public benefits. But once the option is exercised, the resulting demand reduction is an energy product (*ex post*), which should be treated as a private good.

From a systemic perspective, electricity industry provides an essential service with public good attributes, including reliability, environmental, social, and political aspects, all of which must be fully appreciated in comprehensive public policies in ways that align the means and ends. In the retail market, electricity is an essential product but, nevertheless, a private

[9] See FERC (2009).

[10] See Wellinghoff, J. and Morenoff, D. (2007).

[11] In the NOPR, FERC stated:

> We propose that Independent System Operators (ISOs) and Regional Transmission Organizations (RTOs) with tariff provisions permitting demand-response providers to participate as resources in energy markets by reducing consumption of electricity from their expected levels in response to price signals be required to pay to demand-response providers, in all hours, the market price for energy for such reductions.

good.[12] In principle, competition in the wholesale energy market maximizes the gains from trade; this translates into maximizing the sum of consumer surplus and producer surplus, when consumers and producers are allowed to compete on an equal footing for efficient price formation.

The economics of price-responsive demand is well grounded in the engineering-economic reality that electricity generally cannot be stored on an economical basis. The marginal value of electricity consumption and the marginal cost of electricity production could vary significantly from hour to hour. As a result, the real-time market price of electricity, which reflects the marginal cost of production, fluctuates from hour to hour. Price-responsive demand improves the economic efficiency of wholesale electricity markets by discouraging low value energy consumption when the real-time market price is very high and encouraging high value energy consumption when the real-time market price is very low. Price-responsive demand is essential to the efficient pricing of electricity, especially when the supply is scarce. The efficient scarcity pricing rule is to set the price of electricity at the marginal value of consumption, which should be as near the marginal cost of production as possible.

With full price-responsive demand, competitive market prices provide efficient incentives for producing electric services to consumers at the lowest cost in the long term. No further policy intervention is needed. In a hybrid market structure, however, demand response offers public benefits. The lack of price responsiveness during peak periods has been a major concern to policymakers, because electric energy demand and costs tend to concentrate in peak hours.[13] Demand response reduces the ability of suppliers to exercise market power and raise prices above marginal costs, fostering scarcity prices in constrained market areas under tight supply conditions and improving competition in energy markets.[14] Despite the well-recognized

[12] An essential product, like clean air, tends to have high value in consumption, but may have low value in exchange when it is abundant. In contrast, a non-essential product, like diamonds, may have low value in consumption but tends to have a high value in exchange because of scarcity. The basic law of supply and demand dictates that competitive forces will normally drive the market price for any product, whether it is essential or not, toward the marginal value of consumption and the marginal cost of production (although there are well-known exceptions where competitive markets may fail, and in that case, public policy or regulation would be needed).

[13] For example, the 2007 peak load in New England was 26,145 MW. The highest 1% of demand hours accounts for 12% of the peak capacity needs; the 5% highest priced hours account for 12% of the cost of energy. For more general observations, see Faruqui, Hledik, Newell, and Pfeifenberger (2007).

[14] Generally, demand response offers other social benefits such as transmission and distribution network capacity benefits, generation reserve benefits, power quality benefits, environment benefits and renewable energy credits. However, this paper focuses on the energy market within the specified wholesale and retail market structure. We assume that these other benefits are properly priced and are considered outside the scope of this paper.

benefits of price-responsive demand, two market realities that prevent the full realization of price-responsive demand are 1) the lack of advanced metering infrastructure, and 2) the widespread use of fixed uniform retail rates.

The lack of advanced metering infrastructure limits a consumer's ability to access real-time pricing. Evidently, only the size of some large industrial and commercial customers can justify the expense of advanced metering infrastructure, communications, and enabling technologies at this time. Over the past two decades, advances in digital technology have reduced the costs and increased the functionality of smart metering technologies and lowered the entry barrier for price-responsive demand. Recently, these technological barriers have been further reduced through federal and state policies. For example, California's investor-owned utilities are planning to install interval meters for all their customers by the end of 2011. At the federal level, the Energy Independence and Security Act of 2007 provides assistance to promote smart grid activities.

The widespread practice of fixed uniform retail rate has been a strategic barrier that impedes price-responsive demand in wholesale and retail markets for at least two reasons. First, by charging the same rate across a broad range of customers, each independent of individual preferences and load profiles, a fixed uniform retail rate creates cross-subsidies and makes it difficult to recover infrastructure costs without distorting incentives. Customers that consume most of their energy during low-cost, off-peak periods are charged the same price as those who consume most of their energy during high-cost, peak periods. In practical terms, this means users who consume electricity during off-peak periods subsidize peak users. As a result, a fixed uniform retail rate encourages low-value energy consumption when real-time wholesale energy prices are higher than retail rate, creating a tension between those customers who might prefer lower rates and those who prefer more reliable service.

Second, fixed uniform retail rates create disincentives for market participants in the supply-delivery chain to encourage demand response among consumers. Since price-responsive demand generally tends to reduce retail revenue, it creates disincentives for the load-serving entity charging uniform, fixed rates to promote demand response. Decoupling retail revenue from consumption levels has been introduced as a solution to these disincentives. Moreover, the use of fixed uniform retail rates removes the incentive for customers to sign forward contracts that might include participation in wholesale markets through demand bidding. The issue of retail rate reform has been actively debated within individual states. The California Public Utility Commission's landmark decision in 2008 to adopt dynamic pricing as the default rate for all class customers is generally viewed as a positive step toward achieving price-responsive demand.

Ultimately, dynamic pricing that links retail rates to real-time competitive wholesale market energy prices would encourage efficient energy consumption among retail consumers and fulfill the benefits of price-responsive demand. However, the political resistance against dynamic pricing, in large part, is exerted by some large customers with significant peak loads who want to keep their cross-subsidies as long as possible. Initially, demand-response programs would be attractive to those who have significant peak energy consumption or the ability to reduce energy consumption through energy efficiency, while dynamic pricing is attractive to those who have significant off-peak energy consumption or the ability to shift energy consumption to the off-peak period. Therefore, by paying customers to reduce consumption in peak periods, demand-response could lower the barrier to dynamic pricing.[15]

The lack of demand response motivates the use of price cap, which creates the so called “missing money” problem that translates into the need for a capacity market. Paying demand response in the capacity market provides the appropriate source of revenues to encourage the investment in demand response thus lowering the barrier to price-responsive demand. However, demand response resource as an option before it is exercised is different from the resulting demand reduction after the demand response is exercised. In the wholesale energy market, as long as all buyers and sellers own what they sell, demand reduction and generation should be treated similarly as private goods with the same, undifferentiated prices.

Enabling Competition of Differentiated Services

While restructuring has expanded competition in power generation, smart grid opens the possibility of retail competition for differentiated services. Since the 1980s, researchers have begun to envision a future market structure in which retail electric service could be unbundled from local distribution service. While the local distribution company would remain a natural monopoly subject to regulatory oversight, multiple retail firms can compete with each other in providing retail electric services. Retail competition provides the incentive for retail service companies to offer differentiated energy services as enhancements beyond the basic service plans.[16] For example, retail service providers may compete to provide green power to environmentally conscious consumers or blue power to efficiency-conscious consumers (for example, the least-cost bundled services for warmth/coolness, lighting, local communication networks, entertainment, cooking, clothes cleaning/drying, and refrigeration). Reliability differentiated services can be customized for reliability-conscious consumers.[17] For example, some retail service providers may specialize in premium quality power (say,

[15] Dynamic pricing eliminates cross-subsidies, to which peak customers have grown accustomed. The political pressure against rate reform, therefore, is that some want to hang on to their subsidy. On the other hand, for example, plug-in hybrid electricity vehicles that create demand for electric energy during off-peak periods would be an ideal application for dynamic pricing.

[16] See Chao (1989), EPRI Report (2004), EPRI Report (2005).

[17] See EPRI Report (1986).

the platinum or gold power) to certain industrial and commercial customers for whom a reliable electric service is essential. Other retail service providers may offer interruptible service (say, the silver or copper power) to those consumers who are willing to accept service interruptions in exchange for a lower electricity bill. Consumers should have open access to a menu of choices in subscribing the basic and supplemental services. Beyond that, electricity at real-time market prices (or the spot or instant power) should be available as the default service to all customers who want to have unrestricted access to electricity in excess of the amount covered by the basic or supplemental service contracts. However, a key to the realization of these possibilities is the availability of advanced meters and enabling technologies so that contracts for the innovative differentiated service can be designed on the basis of competitive real-time market prices. Essentially, this is a vision of a smart grid world with price-responsive demand.

3. Review of Alternative Approaches

In this section, we review alternative approaches that have been proposed to compensate demand response. We separate these approaches into three groups: the traditional, second-best, and first-best contract-based approaches. Table 1 summarizes the characteristics of these approaches.

Table 1

Alternative Demand-Response Compensation Approaches

	Traditional Approach	Second-best Pricing Approach	First-best Contract-Based Approaches		
			Unbundled Transactions	Buy the Baseline	Demand Subscription
Contract for baseline	No contract	Implicit contract	Implicit contract	Explicit contract	Explicit contract
Who sets customer's baseline	Market operator	Market operator	Market operator or load-serving entity	Load-serving entity or demand-response provider	Retail customer's choice
How much to pay for demand reduction	Wholesale market price	Wholesale market price less retail rate	Wholesale market price	Wholesale market price	Wholesale market price
Who pays for the cost of demand reduction	Load-serving entity and local distribution company	Load-serving entity	No cost allocation is required	No cost allocation is required	No cost allocation is required

3.1 The Traditional Approach

The traditional approach pays a customer for demand reduction at the full wholesale market price, using an administratively determined customer baseline. The customer does not typically pay for energy associated with the customer baseline nor has a financial obligation to purchase the customer baseline amount after receiving demand reduction payment.[18] Most of the initial demand-response programs implemented in the organized markets administered by ISOs and RTOs adopted the traditional approach, and consumers are paid the full wholesale market price or locational marginal price (LMP) for the difference between the actual consumption and a customer baseline estimate.[19] The market operator recovers the demand-response payments by allocating these costs to the market participants using administrative mechanisms.

The traditional approach is inherited from the demand-side management programs in a vertically integrated utility with unspecified ownership or entitlements for the energy from demand reduction. The original idea is attributable to Amory Lovins (1985), who observed that "as long as it is cheaper to save electricity than to make it, both utilities and ratepayers can benefit from properly structured utility financial participation in efficiency."[20] The traditional approach is based on the premise that since a reduction in consumption and an increase in generation produce the same effects in balancing supply and demand, demand reduction should receive a compensation based on the utility's avoided cost, which equals the marginal cost of production. However, in the traditional approach, the ownership of the energy for which the customer is paid pursuant to demand reduction compensation is unspecified.

[18] To measure how much demand is reduced, the traditional approach requires estimation of what a consumer's demand would have been had it not been reduced. This estimated amount is called the customer baseline. Conceptually, the customer baseline is the estimated level of "normal" or counterfactual consumption during the time period against which demand reductions are measured and payments are determined. In practice, the concept must be defined unambiguously so that it can be estimated and verified for the purposes of settlement and billing. The customer baseline is conjectural (i.e., it is not directly observable and is generally estimated from data that represent past customer behavior using statistical estimation methods).

[19] In ISO New England, the existing Day-Ahead Load Response Program pays customers for demand reduction at the full wholesale market price when the wholesale market price exceeds a threshold level. NYISO has similar programs in which customers are paid the full wholesale market price for load reductions in addition to saving the retail rate. PJM had a similar structure until 2008, when it eliminated the extra incentive by deducting the customer's retail rate from the wholesale payment for load reductions.

[20] See Amory B. Lovins (1985).

Economists generally find the above treatment inconsistent with the economic principle of efficient pricing, because it neglects the consumer bill-saving benefits from reduced consumption. This is commonly known as the double-payment issue.[21]

Figure 1 shows the traditional approach in an organized wholesale electricity market administered by an ISO or RTO. The wholesale and retail energy prices are denoted by locational marginal price (LMP_t) and the retail rate (RR), respectively. The customer baseline, i.e., the estimated energy consumption without the demand-response program, is denoted by Q_t; the demand reduction is DR_t; and the actual consumption level is $L_t = Q_t - DR_t$, which varies with time, t.

Figure 1 shows the separate transactions that customer has conducted with the load serving entity (shown in the upper left panel) and with the demand-resource provider (shown in the upper right panel). As a result (shown in the lower panel of Figure 1), the customer pays the load-serving entity (LSE) only for the actual consumption based on metered load and, at the same time, sells demand reduction (DR_t) to the demand-resource provider (DRP). Consequently, the market administrator must reconcile the energy and financial imbalances through demand reconstitution (adding the amount of DR_t) and cost allocation (charging for the amount of LMP_t x DR_t) in the settlement procedure.

[21] See Larry Ruff (2002).

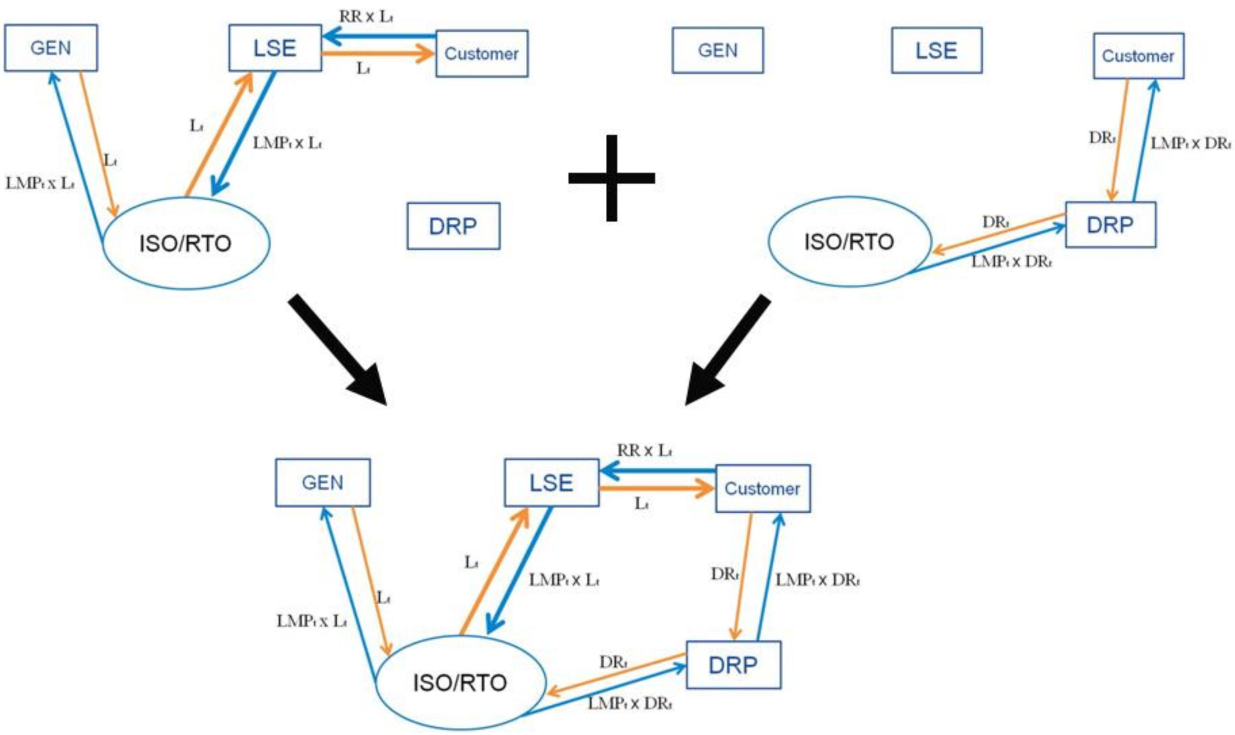

Figure 1: The traditional approach.

The principle of market competition builds on the voluntary exchange of privately owned products and services. Private ownership and property right is the cornerstone of trading. Functional markets generally require that participants are selling what they own. The traditional approach, however, does not require demand-response providers to demonstrate that they clearly and truly own what they are selling. As a result, extending the traditional approach, which was developed in a totally regulated structure, to demand-response programs in a wholesale electricity market environment creates some potentially significant unintended effects. In the following discussion we highlight three types of unintended effects: 1) baseline manipulation, 2) inefficient price formation, and 3) generation relocation and load shifting behind the meters.

Baseline Manipulation

As described above, a customer does not typically pay for energy associated with the administrative baseline nor have a financial obligation to purchase the baseline amount. If customers can consume any quantity at a fixed retail rate and then sell back reductions from the baseline level without a financial commitment to purchase energy at that level, an incentive exists for them to increase their revenues from the demand-response program by increasing their baselines. In the presence of asymmetric information, the administrative customer baseline causes two incentives to increase the baseline: 1) an adverse selection problem and 2) a moral hazard problem.

The adverse selection problem arises from asymmetrical information on customer baselines. That is, since a customer's baseline consumption is not directly observable before participating in a demand response program, customers usually have better information on their baseline consumption levels than the market administrator and can use the information to their advantage in deciding whether or not to participate in demand response programs. Therefore, the program is likely to attract disproportionate participation from customers who anticipate lower consumption for reasons having nothing to do with the demand reduction program. For instance, if last year's consumption is used as the basis for the customer baseline, firms whose production has shrunk since last year are likely to sign up. In this case, consumers could end up paying for demand reduction that would have occurred anyway. At the same time, firms that are entering their high-demand season, or have grown rapidly since last year simply will not sign up.

The moral hazard problem arises from activities that may be undertaken by customers to affect the customer baseline but are difficult to detect. Since the baseline is based on a customer's past consumption, a customer can artificially increase its consumption during "normal" consumption periods to create a higher baseline in order to collect demand reduction payments without actually reducing load. For example, customers with baseload, on-site generation may turn off their generators temporarily to establish an artificially high baseline level of consumption and then turn the on-site generators back on to collect extra payments for what is, otherwise, normal consumption behavior

Inefficient Price Formation

The price formation problem arises from paying too much for demand reduction based on a customer baseline, because it can cause consumers to forgo consumption—even when the value of their consumption exceeds the cost of producing the energy. This happens when consumers receive both bill savings and demand-reduction payment for the same demand reduction, and when the sum of the customer's bill savings and the demand-reduction payment exceeds the cost of production.

Figure 1 shows the effect of double-payment benefits on consumption. Before the demand-response incentive payment, the consumer demand is Q_{peak} when the retail rate is RR. With the incentive payment, a consumer's opportunity cost for demand reduction is the difference between the consumer's willingness to pay (Point D) and the retail rate. When the marginal opportunity cost equals the wholesale real-time market price (or the locational marginal price, LMP), the effect on the customer's consumption is as if the price were the wholesale price plus the retail rate (LMP + RR).[22] As a result, the demand is reduced from the baseline level, Q_{peak}, to Q_1, which is lower than the efficient level of price-responsive demand, Q_0. The incentive is intended to eliminate excessive consumption during peak periods, but it overshoots, resulting in under-consumption in both peak and off-peak periods. As the following analysis shows, the social welfare losses, measured by the shaded area in Figure 1, could outweigh the benefits of price-responsive demand.

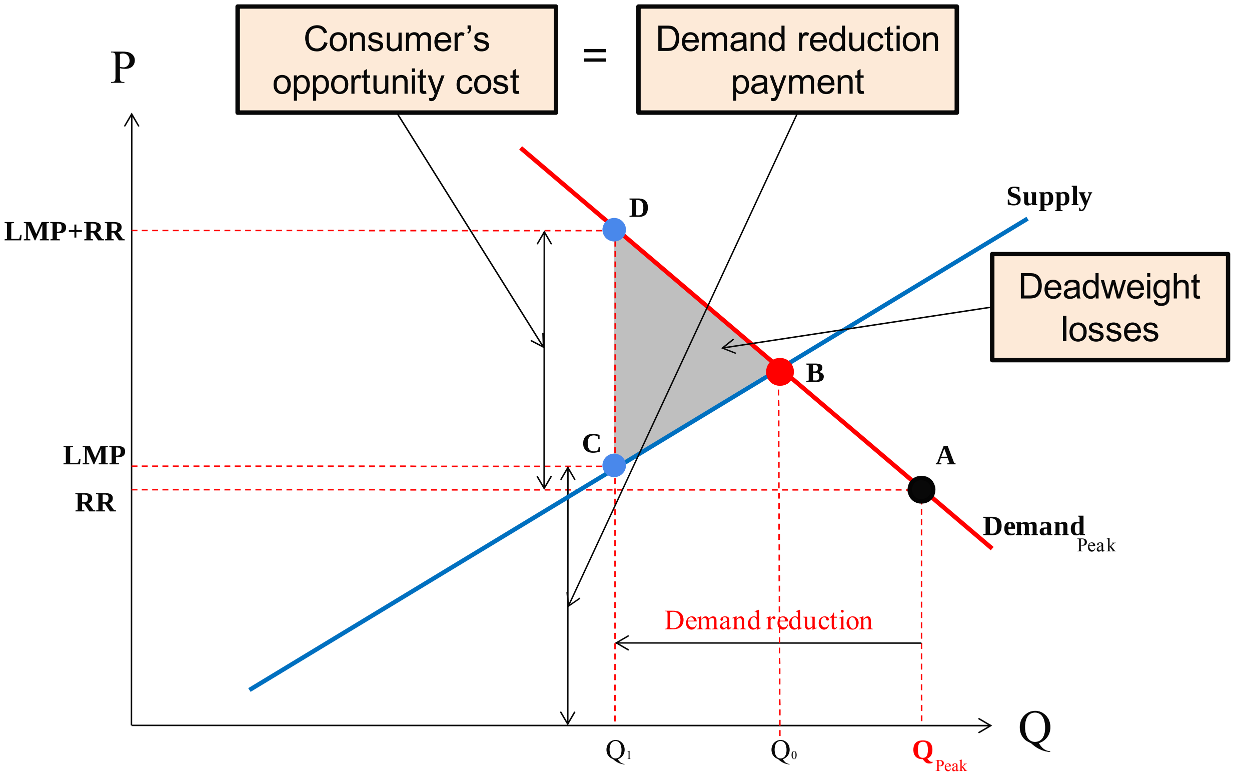

Figure 2 – Double payment incentive causes inefficient price formation

The traditional approach, in effect, creates price differentiation by treating demand reduction more favorably than generation in the wholesale energy market. Demand response is intended to eliminate excessive consumption during peak periods under a fixed retail rate, but the differential treatment "overshoots," resulting in under-consumption, because customers behave (and reduce consumption) as if the price they paid for power were the retail rate plus

[22] For expositional purposes, wholesale real-time market price and locational marginal price (LMP) are interchangeable in this paper.

the wholesale price, because reducing load saves the customer the retail price and additionally triggers an incentive payment equal to the real-time market price.

Below is an example of an inefficient market outcome from price separation between demand-response resources and generators in the wholesale energy market, when risk-free profit can be made simply by putting generators behind the meter.

Suppose that a consumer is willing to pay the retailer up to $150/MWh for energy (in this case, based on the opportunity cost to run an on-site backup generator, including time, materials, and inconvenience).[23] When the retail rate is, for example, $70/MWh, the consumer would be willing to offer demand reduction at $80/MWh. The consumer is willing to offer at this price because the sum of the retail rate savings ($70) and the payment for the reduction ($80 or more) equals or exceeds the cost of running the on-site backup generator ($150).

This would result in the customer operating the backup generator when the cost of producing the energy in the market is lower. For example, once the real-time market price equals or exceeds $80 per MWh, the load reduction offer of $80 would clear in the market and displace the next cheapest generating unit offered at, for example, $100 per MWh. This would result in a net societal cost increase of $50/MWh (*i.e.,* $150 – $100) because the double-payment incentive from the demand response program caused a higher-cost resource to be dispatched – *i.e*., the onsite generator at $150/MWh – while a less-expensive resource – the system generator at $100/MWh – was not utilized. In this example, the market clearing price is artificially depressed – from $100 to $80/MWh, sending incorrect market signals because the market price ignores the cost of running the backup generator at $150/MWh. Society ends up paying a resource cost of $150/MWh, while a substantially less-expensive alternative (*i.e.,* a system generator at $100/MWh) went unused.

The example illustrates that the traditional approach treats demand reductions more favorably than generation. The customer receives $150/MWh in the above example for turning on a backup generator, while the generator that could have produced the energy at $100/MW would have only received $80/MWh and is thus not utilized. For the same reason, the traditional approach also encourages the installation of backup generation, which is an inefficient investment since energy can often be produced more efficiently in the wholesale

[23] The use of on-site generation was used in this example for illustrative purposes only. This example is generally applicable even in situations in which the customer does not have an on-site generator. In that case, a consumer's willingness to pay could be driven by the value derived from consumption. For example, the consumer may be in the business of manufacturing a widget where it is no longer profitable to make the widget if the cost of electricity exceeds $150/MWh. Hence, the manufacturer will cease electricity consumption at prices exceeding $150/MWh.

market. In sum, through the double-payment incentive, a high cost demand response resource can displace a low-cost generator, increasing the resource cost.

Generation Relocation and Load Shifting Behind the Meters

The traditional approach creates incentives for generation relocation and load shifting behind the meters. First, if demand reduction is paid the full LMP without any financial obligation for the baseline, given the opportunity, it would be profitable for generators to relocate from in-front of the meter on the wholesale side to behind a customer's meter to take advantage of the demand response incentives. In the previous example, if the consumer has a contract with a generator with operating cost of $80/MWh and relocate it from the wholesale market to behind the meter, the consumer can offer the energy from the generator as demand reduction back into the wholesale market at the net cost of $10/MWh (i.e., $80 - $70). This offer, if it is cleared in the wholesale market, would reduce the demand and the supply in the wholesale market by an equal amount and thus have no impact on the wholesale market price. The only difference is that the consumer would be paid for the illusory demand reduction created by generation relocation.

Next, if demand-response providers have energy-consuming facilities behind more than one meter, the traditional approach can present incentives to shift load behind those meters to create demand reduction that is illusory. In the traditional approach, the ownership of demand resources in the wholesale market tends to be obscured with the use of an administrative customer baseline. Under the traditional approach, the consumer does not buy the baseline nor have an obligation to pay for it. This enables unproductive load-shifting activities behind multiple meters that make it appear that energy sold on the wholesale energy market was being created when it was not. It is instructive to consider a simple example.

Suppose that a manufacturing company owns two identical facilities behind separate meters in the same RTO region. At the beginning, both facilities consume identical amounts of electricity at 80 MW/hour, when they run at the usual 80% of capacity. Before participating in the demand-response program, both facilities have the same customer baseline of 80 MW/hour. If both facilities participate in a demand-response program as separate assets, and each facility is paid the real-time market price for the reduced consumption below the customer baseline, then the company could implement a load-shifting strategy between the two facilities as illustrated in Table 2.

Table 2
Load Shifting Behind Multiple Meters

	Day One			Day Two		
	Customer Baseline (MW)	**Electricity Consumption (MW)**	**Demand Reduction (MW)**	**Customer Baseline (MW)**	**Electricity Consumption (MW)**	**Demand Reduction (MW)**
Facility 1	80	100	—	82	60	22
Facility 2	80	60	20	80	100	—
Total		160	20		160	22

On Day One, the company could control the resources (e.g., labor and materials) at the two facilities so that Facility 1 runs at 100% capacity while Facility 2 runs at 60% capacity. On this day, these facilities will consume, respectively, 100 MW and 60 MW/hour for a total of 160 MW/hour. While Facility 1 consumes 20 MW above the current baseline and will pay for the extra energy consumption at the regular fixed retail rate, but this increased expense is offset by savings at Facility 2. Moreover, since Facility 2 consumes 20 MW below the baseline, it can sell this amount as demand reduction into the wholesale energy market and be paid the real-time market price for the demand reduction. Based on the methodology currently used to calculate customer baselines described in the ISO or RTO tariff, the baseline for Facility 1 would be raised to 82 MW while the baseline for Facility 2 will remain the same at 80 MW for Day Two. [24]

On Day Two, the allocation could be reversed so that Facility 1 runs at 60% capacity while Facility 2 runs at 100%. As a result, the two facilities will consume the same amount of electricity as on Day One (100 MW and 60 MW, respectively), except that their positions are reversed. Now Facility 2 will be billed for 20 MW/hour at the fixed retail rate, which is offset by the savings at Facility 1, but Facility 1 can sell 22 MW of the amount of its consumption below the baseline as demand reduction and be paid the real-time market price as extra benefits.

The above load-shifting strategy can be repeated between the two facilities every day. On alternating days, one of the facilities is running at 60%, so the electricity consumption at that facility would be 60 MWh, which is 20 or 22 MW lower than the baseline. Therefore, the company could submit demand-response bids for the facility in the wholesale electricity market and get paid for the 20 or 22 MW of demand-reduction payment—even though the total electricity consumption stays at 160 MWh/day. In this case, the demand reduction created by the load-shifting strategy is illusory, and not genuine. Other customers bear the

[24] The baselines at both facilities may increase such that the calculated load reduction can grow over time. This is occurs because baselines in ISO-NE are currently calculated using averages that exclude days when the demand reduction offer clears.

burden of payments for the illusory demand reduction, while the manufacturing company receives the payments at virtually zero opportunity costs.

The above example suggests that even without ostensibly unlawful manipulation or fraudulent activities, illusory demand reduction can be created in the wholesale market when there is no real reduction in electricity consumption. This is possible primarily because first, customers have no financial obligation to pay for the customer baseline and second, customers can still buy any amount above the baseline at the fixed retail rate. In the above example, if the manufacturing company has the financial obligation to pay for its customer baseline, then the procurement cost of the energy for demand reduction in one facility will be comparable with the cost for the increased consumption in the other facility. Moreover, if the firm has to pay LMP for any amount of energy above the baseline, the load-shifting strategy will no longer have value to the firm.

To sum up, the traditional approach treats demand reduction as if consumers own what they are selling when, in fact, they have not exercised the quantity call option to establish contractually enforceable rights and obligations for the customer baseline. This explains the moral hazard and adverse selection problems that tend to exaggerate the customer baseline. In addition, the double-payment incentives result in excessive supply of demand reduction that depresses prices to inefficient levels in the wholesale market. Lastly, the different treatment between demand reduction and generation along with the cost allocation mechanism, in effect, creates price differentiation among consumers and generators in a basically homogeneous energy market. In most market settings, price differentiation normally arises from conditions of market segmentation. Absent market segmentation, price differentiation of a homogeneous product is susceptible to exploitative gaming strategies that may undermine market integrity.[25] More sophisticated measurement and verification methods for estimating the customer baseline can help, but will not solve the problem that is fundamentally created by the absence of ownership of the customer baseline. Essentially, allowing the demand-response providers to sell what they do not own is a fundamental market design flaw that may distort economic efficiency and compromise market integrity.

In the following sections, we examine alternative remedies that have been proposed.

[25] Two historical precedents highlight the importance of proper alignment between pricing structure and market conditions. First, as indicated in Hogan (2010), in 1997, the PJM electricity market initially had a flawed design that precludes efficient congestion management and nodal price separation, despite the reality that market separation might occur due to transmission congestion. Second, as is widely recognized, the California electricity market design initially adopted a flawed market design that did not allow price separation within a load zone creating opportunities for some well-known gaming strategies that contributed to the 2000-2001 California electricity crisis.

3.2 The Second-Best Pricing Approach

The initial demand-response programs in the organized electricity markets mostly adopted the traditional approach of paying demand reduction at the full wholesale price. Given the concerns of distorted incentives, stakeholders in the PJM and New England regions have discussed proposals that set the demand-response payment at the difference between the wholesale price and the retail rate.

As will be shown below, the optimal level of demand-response payment in a hybrid market structure equals the difference between the wholesale price and the retail rate. An advantage of this approach is that it can be implemented within the jurisdiction of wholesale market depending on regional variations in retail tariff. However, this approach represents a compromised, second-best solution because while it reduces over-consumption when the wholesale spot price is higher than the retail rate, it does not address the under-consumption inefficiency during the off-peak period when the wholesale price is lower than the fixed retail rate.

Consumers participating in a demand-response program can submit bids in the wholesale market. Setting the optimal demand-response payment must take into account the incentives that consumers face. In the following examples, we show that if the demand-response incentive equals the real-time market price minus the retail rate (LMP – RR), the optimal bids should reflect the incremental values of service, denoted by VOS; and that when the consumer demand bid sets LMP, it should reflect the true marginal value of consumption.

Let's consider two cases. In the first case, the consumer bids an amount Bid > VOS. If the bid is cleared (i.e., Bid < LMP), then the consumer will be paid (LMP − RR) for the demand reduction. Since (LMP − RR) > (Bid − RR) > (VOS − RR), the consumer will get the same result without overbidding. But, if (Bid > LMP) and the bid is rejected, the consumer will keep the consumer surplus (VOS − RR), which could be less than the incentive payment (LMP − RR) when Bid > LMP > VOS. Therefore, over-bidding never pays off.

In the second case, the consumer's bid is less than the true value (Bid < VOS). If the bid is cleared, (i.e., Bid < LMP) the consumer will be paid (LMP − RR), which may be less than the consumer surplus (VOS − RR) when VOS > LMP > Bid. Therefore, the consumer would be better off by bidding VOS truthfully and keeping the higher consumer surplus. On the other hand, if the bid is rejected, the consumer will keep the consumer surplus (VOS − RR), which is higher than the demand-response payment (LMP − RR) because VOS > Bid > LMP.

The second case shows that underbidding does not pay off either. As a result, the incentive for the consumer is to reflect the value of service VOS in the demand bid. If the demand bid sets the wholesale market clearing price, then the price should reflect the marginal value of consumption. This result is consistent with the marginal cost pricing principle. The effect of the demand-response payment is to induce consumption at a level as if the consumer were paying the wholesale price. This demonstrates that the wholesale energy prices provide efficient price signals.

In the special case when the retail rate *equals* the wholesale price, the optimal demand-response payment would be zero. Therefore, for consumers on dynamic retail pricing, there is no longer any reason to pay them for demand reduction.[26] In contrast, if the consumer has a take-or-pay long-term contract, then the optimal demand-response incentive payment would be the full wholesale price, because the effective retail rate under contract is zero. If the demand-response payment is set above the optimal level, it is likely to attract too many consumers to the demand-response program and impede the prospects of dynamic retail pricing and long-term contracting.

The apparent differences in the payment rule for demand response and generating resources is because absent ownership of the baseline, demand response resource remains an option that has not been fully exercised; demand response resource with double payment incentives is different from one with an established baseline. An important advantage of the second-best pricing approach is that it corrects the double-payment incentives and thus restores efficient price formation by directly deducting the retail rate from the wholesale energy payment to demand-response providers. In effect, the market operator is administering an implicit contract for a consumer that procures energy at the retail energy rate and sells it at the wholesale market price. This approach does not eliminate but reduces the severity of the problems of baseline manipulation and load shifting behind the meters , because no demand-response payments would be made when wholesale prices are at or below the retail price.

3.3 The Contracted-based Approaches

In this section, we review three contract-based approaches that have been proposed to fix the problems associated with the use of administratively-determined customer baseline, 1) unbundled transactions, 2) buy the baseline, and 3) demand subscription.

Unbundled Transactions

The unbundled transactions approach proposed by William Hogan provides a simple and theoretically elegant framework that addresses the double payment issue. The basic idea of

[26] This is consistent with the PJM market monitor's assessment that customers already paying the wholesale price should receive no payment for reducing usage. See "MMU White Paper: PJM Demand Response Program," December 4, 2007, page 7 http://www.pjm.com/committees/drsc/downloads/20080509-item-01-whitepaper.pdf.

the unbundled transactions approach is to separate the demand-side transaction into two different market transactions: 1) energy procurement in the retail market, and 2) energy sale in the wholesale market. The demand reduction could still be estimated as the difference between an estimated customer baseline and the actual consumption. Unlike in the traditional approach, however, the consumers are imputed an implicit contract for procuring the amount of reduced demand from their retail service provider or load-serving entity to obtain the right to sell it. With the implicit contract, the consumer would be deemed to have purchased this amount so that, separately, the customer could sell the same quantity into the wholesale energy market and be paid the market price, as with other energy sales from the generators. Hogan (2009) explains the merits of unbundled transactions succinctly, as follows:[27]

> The very essence of demand-response programs is that, but for the actions of the customer, the notional electricity would be purchased and consumed. It is this deemed energy purchase that is presumably being sold to the system operator. The separation of the net transaction into the purchase and the sale clarifies the underlying economics and makes it clear that putting demand response on a level playing field with generation must require that the net treatment of demand response be the same as the net treatment of the purchase and sale of the notional electricity.

In the unbundled transactions approach, the customer pays for the total of the actual consumption and the demand reduction. This has two effects on implementation. On the one hand, the wholesale market settlement system is in balance. Therefore, there is no need for the difficult and contentious load-reconstitution and cost-allocation procedures. On the other hand, it requires the load-serving entity to adopt a new system of billing customers for demand reduction (the amount of energy that they did not consume). In effect, the load-serving entity is required to implement an accounting system for the customer baseline. This is a nontrivial change from the traditional practice of billing customers only for the actual consumption that is metered.

Buy the Baseline

A demand-response provider could be required to purchase its expected amount of energy consumption in advance and schedule it in the day-ahead energy market in order to be paid the wholesale price in the real-time energy market. This approach is referred to as “buy the baseline” approach.

[27] See Hogan (2009) and Cicchetti and Hogan (1989).

Under this approach, the consumer can buy or schedule the baseline in the Day-Ahead Energy Market and then sell any demand reduction subsequently in the Real-Time Energy Market. This can be interpreted as a special realization of the unbundled transactions approach, in which the two separate transactions are settled separately in the day-ahead and real-time markets. The principal characteristic of this approach is the elimination of the need to use an administrative customer baseline. Bushnell, Hobbs, and Wolak (2009) articulate the advantages of such an approach as follows[28]:

> Final consumers can schedule a given level of consumption in the day-ahead market and then sell day-ahead ancillary service capacity or real-time energy reductions relative to this day-ahead schedule in the real-time market. Similarly suppliers can schedule from their generation units in the day market and then sell day-ahead ancillary services or additional energy in the real-time market beyond that day-ahead energy schedule. This "buy your baseline" approach to selling demand reductions in a subsequent market ensures that retailers and curtailment service providers (CSPs) face the full financial consequences of their baseline choice in the same way that suppliers face the full financial consequences of their final energy schedules in the real-time market.

In the wholesale electricity market in New England, this buy-the-baseline approach was called the hybrid approach because it implements the unbundled transactions on the demand side (in the day-ahead market) and the supply side (in the real-time market). This approach was initially considered in the stakeholder process but was dropped, given greater stakeholder support for the second-best approach described in section 3.2.[29]

Basically, the buy-the-baseline approach requires demand-response providers to establish customer baselines and settle the costs of demand response in the day-ahead market. Therefore, the buy-the-baseline approach obviates the need for the decisions on how to allocate costs of demand response to market participants.

Buy-the-baseline can be viewed as a wholesale market approach that allows the retail customer to contract on a forward basis (e.g., with a demand-response provider) a specific amount of energy and take ownership of the amount contracted. Once the customer baseline level is established and scheduled in the wholesale day-ahead market, demand reductions can be offered into the wholesale real-time market and compete with supply resources on a comparable basis. This improves the economic efficiency of wholesale electricity markets by discouraging low-value energy consumption when wholesale energy prices are very high.

[28] See Bushnell, Hobbs, and Wolak (2009).

[29] See ISO-NE (2009).

This should lower demand during times of system peak and near system peak. As a result, the market clearing price would send efficient price signals that reflect the scarcity value of resources during the shortage hours and achieve the efficient level of price-responsive demand.

Demand Subscription

In its simplest form, demand subscription allows a consumer to contract a specific amount of electricity at a fixed price before actual consumption.[30] The differences between the actual consumption and the subscription level accumulated over a pricing period can be settled at a default rate. A demand-response program allows the unused amount of subscription to be sold as demand reduction in the wholesale energy market. Demand subscription can be interpreted as an alternative realization of the unbundled transactions approach as well as an extension of the buy-the-baseline approach in the retail market. Its principal characteristic is that each final consumer or end-user can acquire ownership of the baseline on the basis of self-selection. This minimizes the incentives for gaming.

Demand subscription addresses the root cause of the customer baseline problem that it lacks ownership. Demand subscription establishes the entitlement and property rights of the energy that a consumer could later sell as demand reduction. Thus, the customer could legitimately "sell-back" the unused energy below the subscription level. Once the customer baseline level is established by demand subscription, demand reductions can be offered into the wholesale market. In this way, every customer would have an incentive to make an offer that reflects the true value or opportunity cost. As a result, the wholesale market would send the right price signals.

Demand subscription is a retail rate design approach that builds on a two-settlement transaction system in which the basic electric service is unbundled into a standard service and a default service. The standard service is settled ex ante, and the default service is settled ex post. A key advantage of service unbundling is to give customer choices, an opportunity to reduce the cost of the price-quantity insurances that are bundled in the full requirement service.[31] Each customer can choose between standard service contracts for preferred hedges against price volatility and the default rate for the unlimited supply of last-resort. A customer can get refund for any unused amounts of energy subscribed at the default rate or sell them into the wholesale energy market as demand-response resources.

[30] See Chao (2010).

[31] See EPRI (1986).

In more general forms, multi-level demand subscription provides a mechanism that allows consumers to choose differentiated services based on service priority or load profile.[32] The standard service may include a menu of differentiated service options with different price thresholds. In principle, a consumer could choose various amounts of energy with a set of subscription prices and threshold prices based on individual preferences. Each service option is a call option that can be interrupted when the real-time market price exceeds a threshold price. Essentially, the demand-subscription approach unbundles the full-requirement electric service, transforming an unlimited quantity call option into a spectrum of limited call options based on the principle of consumer self-selection. [33] Demand subscription establishes a platform that allows third-party service providers to compete in value-added service offerings and innovative products beyond the standard and default services in the basic service plan. The general incentive approach based on customer self-selection enables more flexible pricing structure, which reduces the consumer's risk.

The default rate may be linked to the real-time market price. Demand subscription essentially implements real-time pricing if the default rate equals the real-time market price. In principle, customers can procure, at the default rate, any amount of energy above the subscription level which naturally defines a two-sided baseline, as in two-part real-time pricing. For example, on one hand, if the electricity consumption is below the subscription level during peak hours, the demand reduction can be sold into the wholesale market when the real-time market price is high. On the other hand, the consumer can also increase the electricity consumption above the subscription level during off-peak hours or shift load from peak to off-peak hours, and pay the wholesale market price when the price is low. As a result, demand subscription encourages consumers to opt in for real-time pricing fostering the full potential of price-responsive demand.

Once the customer baseline level is established by demand subscription, demand resources can be offered into the wholesale market on a similar footing to supply resources. In this way, the wholesale market attracts and retains those customers willing to reduce electricity use during shortage conditions in competition with generators. This improves the efficiency of price formation in wholesale electricity markets by discouraging low-value energy consumption when wholesale energy prices are very high. This should lower demand during times of system peak and near system peak.

For example, a load-serving entity or a retail service provider/aggregator could make an offer on behalf of a customer in the energy market based on the customer's expected value of

[32] See Chao, Oren, Smith and Wilson (1986), and Chao and Wilson (1987).

[33] See EPRI (1986). It shows that multilevel demand subscription allows consumers to unbundle retail services and increase the range of consumer choices.

service (VOS). During peak hours, demand resources would be dispatched in merit order based on the offers. When the resource is dispatched, the customer would lose the value of service. Therefore, if the wholesale energy price were greater than the offer price of the consumer's VOS, the demand resource would be dispatched. The demand reduction would be compensated at an amount equal to the wholesale energy price, which will be equal to or greater than VOS.

Alternatively, if the wholesale energy price were less than the VOS, the demand resource would not be dispatched, and the customer would keep its value of service. Therefore, a demand resource would be dispatched, or a demand reduction requested, during the shortage hours only if the wholesale energy price is large enough to compensate for the losses in consumer value (VOS). In this way, every customer would have an incentive to make an offer that reflected the true value of service. As a result, the market clearing price would send efficient price signals that reflected the scarcity value of resources during the shortage hours and achieve the efficient level of price-responsive demand.

Practically, the transition to a two-settlement system can be accomplished in a voluntary manner, assuring that no consumers would be disadvantaged. Initially, each consumer is allocated a customer baseline based on the load profile expected under the cost-of-service fixed retail rate. The initial allocation ensures that no customer will pay more for the same service under the fixed retail rate. If some customers find it better off to reduce the demand during the peak period when the wholesale market price is higher than the values that they derived from electricity service, they are compensated at the wholesale market price. No other consumers should be worse off because the demand reduction will lower the wholesale market price and the total resource cost. Further cost savings can be obtained by introducing priority service as an efficient rationing scheme for curtailing excess demand. As a result, the average cost of service for the remaining consumers will be lower. Priority service dominates an indiscriminate, random rationing scheme in the sense that if the cost savings are refunded equally to all customers in proportion to their energy consumption, then every customer can be made better off or, at least, not worse off.[34] In the long-term, implementing demand response using the demand subscription approach with a two-settlement system can achieve similar social welfare and efficiency results to real-time pricing but with different allocative impacts depending on the range of tariffs available to consumers. Consumers who prefer greater price hedge may increase the level of subscription, and others who are risk neutral may opt for dynamic pricing in the default rate.

4. An Economic Evaluation of Alternative Approaches

[34] See Chao and Wilson (1986) for a theoretical proof.

In this section, we illustrate the economic benefits of demand response using a simple example. For expository purposes, we assume:

1. The wholesale and retail markets are perfectly competitive.
2. The wholesale market price equals the marginal cost of supply.
3. The profit margin for retail service sales is zero as a result of free entry.

Therefore, the retail price equals the expected demand-weighted average of wholesale prices.

Figure 3 illustrates the economic benefits of price-responsive demand measured by the increase in social surplus when customers change consumption levels in response to wholesale market prices.[35] Price-responsive demand improves economic efficiency in both peak and off-peak periods. During the peak period, demand reduction improves efficiency because it cuts consumption that is valued less than the cost of production. During the off-peak period, the real-time market price is less than the uniform price, so increased demand during such times improves economic efficiency because the marginal value of consumption is greater than the marginal cost of production. Price-responsive demand avoids over-consuming when wholesale prices exceed the uniform retail rate or under-consuming when wholesale prices are below the uniform retail rate. The shaded areas in Figure 4 show that the social surplus is increased by eliminating the deadweight losses.

[35] Social surplus is the sum of consumers' surplus and producers' surplus (where the consumer surplus is the amount that consumers benefit by buying electricity at a price that is less than they would be willing to pay, and the producer surplus is the amount that producers benefit by selling at a market price that is higher than they would be willing to sell). Social surplus measures the net social benefit achieved by market trading between buyers and sellers. Therefore, social surplus offers a metric by which market performance can be measured.

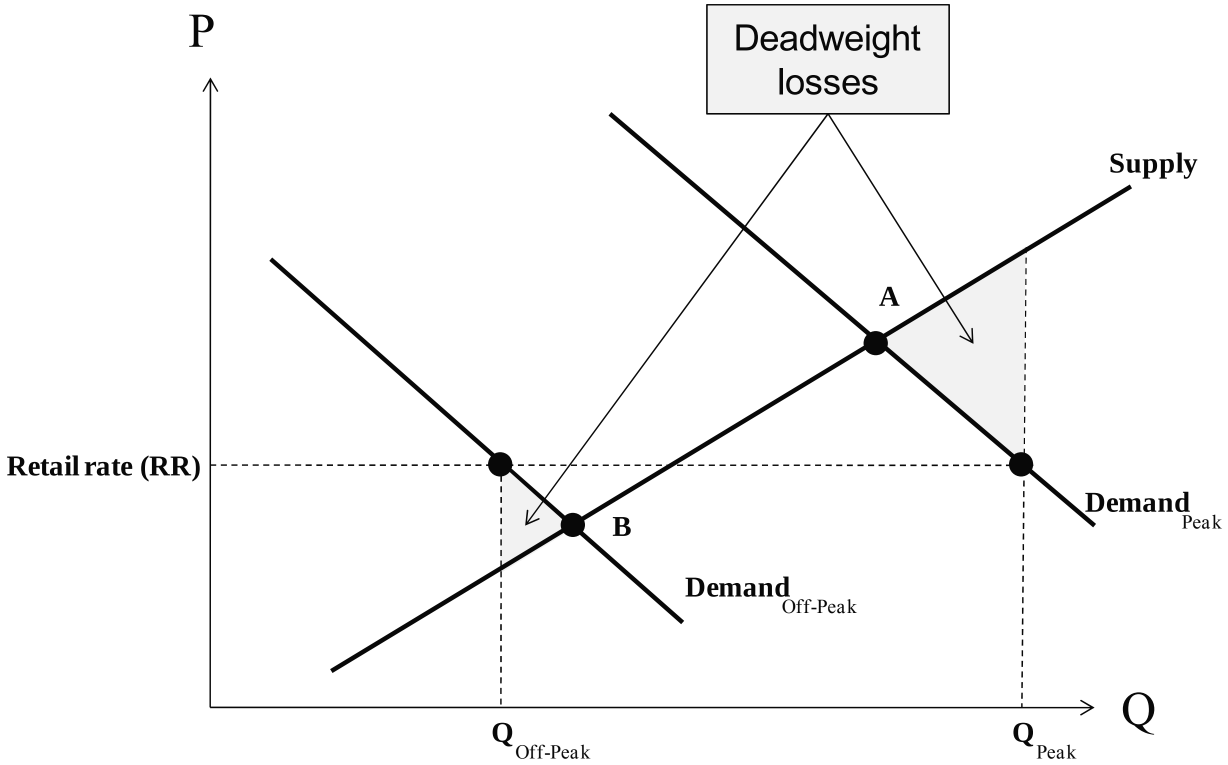

Figure 3: Economic benefits of price-responsive demand.

4.1 Short-Run Analysis[36]

We consider a power system that has a peak load period for 25% of the hours what hours? and an off-peak period for 75% of the hours. The hourly demand functions (measured in MW) for electric energy in these periods varies linearly with the price paid by the consumer as follows:

Peak hour: $$D_{Peak}(P) = 22{,}000 - 20 \times P \quad (1)$$

Off-peak hour: $$D_{Off-Peak}(P) = 11{,}000 - 10 \times P \quad (2)$$

The short-run supply function (measured in MW) is the same for both peak and off-peak periods as follows:

[36] The analysis uses an economic model with a hybrid market structure that is summarized in a separate paper. See Chao (2009). For an application to the ISO New England Market, see ISO-NE (2009).

$$\text{All hours:} \qquad S(P) = 4{,}000 + 100 \times P \qquad (3)$$

In the following, we consider five cases.

Case 1: Price-Insensitive Demand: All consumers are under a fixed-price retail rate.
Case 2: Price-Responsive Demand: All consumers are under real-time pricing.
Case 3: Traditional Approach:[37] The demand reduction is paid at the real-time market price, or LMP, which equals the all-inclusive wholesale price.
Case 4: Second-Best Approach: The demand reduction is paid LMP – RR, if the LMP is higher than the RR.
Case 5: First-Best Approach:[38] Each consumer subscribes for a specific level of standard service. The demand reduction from the subscribed level is paid LMP.

Cases 1 and 2 represent two possible benchmarks. In Case 1, consumers are charged a fixed retail rate and the demand is price-insensitive. This case roughly represents the current situation. In Case 2, all retail consumers pay the real-time market price. This case represents the ideal outcome. Cases 3 to 5 span the range of alternative approaches that have been considered for demand response. In these cases, we assume full participation for expositional purposes, and show how the efficiency of demand response depends on the economic incentives.

Case 1: Price-Insensitive Demand: All consumers are under fixed uniform pricing

In Case 1, demand is insensitive to wholesale spot prices. This case represents the status quo in which consumers do not participate in demand-response programs. This serves as a reference case against which the performance of other approaches is evaluated. As shown in Figure 4, the consumption levels are set at the points, Q_{Peak} and $Q_{Off\text{-}Peak}$, on the demand functions where the price equals the fixed-price uniform retail rate (RR). This causes over-consumption in peak periods (at Q_{Peak}) and under-consumption in off-peak periods (at $Q_{Off\text{-}Peak}$), each leading to deadweight loss (shaded gray).

[37] In New England, the existing Day-Ahead Load Response Program pays customers for demand reduction at the full LMP when the LMP exceeds a threshold level. NYISO administers similar programs in which customers are paid the full LMP for load reductions in addition to saving the retail rate. PJM had a similar structure until 2008, when it eliminated the extra incentive by deducting the customer's retail rate from the wholesale payment for load reductions.

[38] For the economic efficiency viewpoint, all of the first-best solutions (i.e., demand subscription, buy the baseline, and unbundled transaction), are equivalent. For expository purposes, the case illustrated in this example is based on retail demand subscription.

Table 3 displays results of Case 1 and Case 2. In Case 1, the retail rate equals $100/MWh. From the demand functions in Cases 1 and 2, this price implies that the peak demand is 20,000 MW, and the off-peak demand is 10,000 MW. From the supply function in (3), these levels of output imply that the wholesale market price equals $160/MWh and $60/MWh, respectively, during peak and off-peak hours. Note that the retail rate equals the demand-weighted average of wholesale prices.

Table 3 Price insensitive demand vs. Price-responsive demand

	Case 1. Price-insensitive Demand	Case 2. Price-responsive Demand
Peak hour (6 hours)		
Demand (MW)	20,000	19,000
Wholesale price ($/MWh)	$160.00	$150.00
Retail price ($/MWh)	$100.00	$150.00
∆ Consumer surplus from Case 1 ($Million/Year)	-	($2,135)
∆ Producer surplus from Case 1 ($Million/Year)	-	($427)
Off-Peak hour (18 hours)		
Demand (MW)	10,000	10,364
Wholesale price ($/MWh)	$60.00	$63.64
Retail price ($/MWh)	$100.00	$63.64
∆ Consumer surplus from Case 1 ($Million/Year)	-	$2,433
∆ Producer surplus from Case 1 ($Million/Year)	-	$243
All Hours		
Energy consumption (GWh/Year)	109,500	109,699
∆ Consumer surplus from Case 1 ($Million/Year)	-	$297
∆ Producer surplus from Case 1 ($Million/Year)	-	($184)
∆ Social surplus from Case 1 ($Million/Year)	-	$113

Case 2: Price-Responsive Demand: All consumers are under real-time pricing

In Case 2, under real-time pricing, all consumers pay real-time retail prices that reflect the real-time market prices in the wholesale market, and demand is responsive to real-time market prices. Thus, as shown in Figure 4, consumption occurs at the points (A and B) along the demand curve that intersects the supply curve, i.e., according to the real-time market price in the wholesale market. This achieves the optimum allocation with no deadweight loss.

Table 3 shows that the energy demand during the peak hours is reduced by 5% to 19,000 MW when the energy price increases from $100/MWh to $150/MWh. In contrast, the level of demand during off-peak hours rises by 3.6% to 10,364 MW as the energy price drops from $100/MWh to $63.64/MWh. Overall, the total level of demand is increased by 0.18%.

The net social benefit of moving from a fixed retail rate to real-time pricing is $113 million per year, which is the sum of gains in consumer surplus of $297 million/year and losses in producer surplus of $184 million per year. The losses for the producers would reduce new investment in generation capacity. The reduction in peak demand should reduce the investment required in the long term.

The social benefits and costs are distributed asymmetrically between peak and off-peak hours. During peak hours, both consumers and producers incur net losses. Consumers are paying the real-time market price of $150/MWh, which is higher than the fixed retail rate of $100/MWh, and the producers are paid the real-time market price at $150/MWh, which is lower than the previous wholesale price of $160/MWh. During off-peak hours, both consumers and producers are better off with real-time pricing, because the price is lower for consumers and the demand is higher for the producers.

The asymmetrical distribution reflects the cross-subsidies in the fixed uniform retail rate. Many large industrial and commercial customers with significant peak load are unwilling to switch to real-time pricing because the benefits of switching to the lower real-time market prices during off-peak hours are not large enough to cover the loss of cross-subsidies plus the cost of infrastructure investment, even though it may be economically efficient to do so. As shown in Table 3, real-time pricing reduces the consumer surplus during the peak hours by $2,135 million per year, but it increases the consumer surplus during the off-peak hours by $2,433 million per year, with a net benefit of $297 million per year.Consumers with significant peak consumption would be worse off unless they gain sufficient benefits from the low prices during the off-peak period to offset the losses during the peak period.

Demand-response programs paying consumers to forgo consumption during peak periods can be interpreted as a risk-sharing mechanism that mitigates the risk caused by losing cross-subsidies.[39] When some consumers reduce demand during a peak, high-priced period, the other consumers reap the spillover benefits of the lower prices that reduce the cost of service. Moreover, lower peak prices would reduce cross-subsidies and lower the barrier to price-responsive demand. Therefore, in principle, demand-response programs could jump-start the transition to price-responsive demand.

[39] For a risk management perspective, see Chao, Oren, and Wilson (2005).

Table 4 displays results of Cases 3 to 5, comparing alternative approaches for demand-reduction payment.

Case 3: Traditional Approach: The demand reduction is paid at the real-time market price

This case corresponds roughly to the current demand-response programs in which participants can continue paying a uniform retail rate for only metered load, and are also paid the wholesale real-time market price for load reductions from a higher customer baseline level. In this case, customers consume (and reduce) as if the price they paid for energy were the retail rate plus the real-time market price, because reducing load saves the customer the retail rate and additionally triggers an incentive payment equal to the wholesale real-time market price. This incentive is intended to eliminate the excessive consumption during peak periods under the uniform retail rate, but it overshoots, resulting in under-consumption in every period.

In this case, the demand-reduction payments are allocated to the load-serving entities, and retail rates are adjusted to include this cost in all customers' bills. Because of the under-consumption, the wholesale market clears at a lower point on the supply curve. As a result, the retail rate is still lower than that in Case 1. However, this price effect would likely diminish over time as suppliers adjust their investment/retirement decisions in response to lower wholesale prices.

Case 4: Second-Best Approach: The demand reduction is paid LMP - RR, whenever the LMP is higher than the RR.

As in Case 3, demand-response program participants can continue paying a fixed retail rate for only metered load and will receive demand-reduction payments. But in Case 4, the demand-reduction payment for load reductions is based on the wholesale real-time market price (LMP) minus the retail rate (RR) when the wholesale price is higher than the retail rate. In Case 4, the consumption level is determined as if all demand is responsive to wholesale real-time market prices whenever the real-time market price exceeds the retail rate. As in Case 2, this eliminates inefficient over-consumption during high-priced hours, but without overshooting as in Case 3. However, this approach does not directly address under-consumption during low-priced hours. There is, however, a slight reduction in under-consumption inefficiencies because reduced wholesale prices and consumption during high-priced hours lower the retail rate needed to achieve revenue adequacy over the whole year. The lower retail rate slightly increases consumption in the low-priced hours in the wholesale energy market when customers would not be price-responsive.

Case 5: Contract-Based Approach: Each consumer subscribes a specific level of baseline service. The demand reduction is paid LMP.

Demand subscription facilitates integration of demand response in the wholesale market. Once the customer baseline level is established through demand subscription, demand reductions can be offered into the wholesale market on a similar footing to supply resources. In this way, the wholesale market attracts and retains those customers willing to reduce electricity use during peak hours in competition with traditional generators.

To illustrate, in Case 5, we consider a simple example in which all consumers participate in demand subscription with a simple option: each consumer contracts a fixed level of consumption and can exercise the option to curtail the demand when the wholesale price exceeds the retail rate, and the amount of demand reduction. That is, a demand-response program allows the consumer to sell back any unused amount below the subscribed amount when the wholesale price is greater than the retail rate. It also allows the consumer to buy any amount above the subscribed level at the wholesale market price when the price is low. Therefore, demand subscription naturally defines a contract-based "two-sided" customer baseline. This provides efficient incentives for consumer to shift consumption from peak to off-peak hours, reducing consumption during high-price periods and increasing consumption during low-price periods. As a result, demand subscription achieves efficient price-responsive demand in much the same way as real-time pricing.

In Case 5, we assume that each consumer may choose between the standard firm service (associated with a sufficiently high threshold price so that it is practically never interrupted) and an interruptible service (associated with a threshold price equal to the retail rate). Therefore, when the wholesale price is greater than the retail rate, the interruptible service may be interrupted, or equivalently, the consumer can sell back any unused amount of subscription as an interruptible service. When the real-time market price falls below the retail rate, the consumer has the option to buy any amount of energy above the subscribed level at the wholesale market price.

The most surprising and important insight is that Case 3 (paying for demand reduction at the full wholesale price with double payment benefits) is the least-efficient approach, substantially less efficient even than fixed-price rate. Whereas the optimal level of consumption (Case 2) is where the demand responds to the real-time market price, demand in Case 3 responds to a net price signal that always exceeds the real-time market price. For example, if the retail rate is $95/MWh and the real-time market price is $134/MWh, a customer would be able to earn $29/MWh by using a $200 generator (or forgoing the value of consumption worth $200) to reduce its net metered load, even though generating for $200 when the real-time market price is $134 is inefficient (i.e., using a $200 generator or forgoing the consumption worth $200 when a $134 generator was available is clearly an inefficient use of society's resources).

It is instructive to compare demand subscription (Case 5) with price-responsive demand (Case 2). Both yield the same wholesale price ($150/MWh) during peak hours and $63.64/MWh during off-peak hours, and with the same social surplus increase of $113 million per year from Case 1. However, the two cases differ in the way consumer benefits are distributed between peak and off-peak hours. Switching from fixed retail rate to real-time pricing, consumers gain an increase in consumer surplus of $2,433 million per year during off-peak hours when the price drops from $100/MWh to $63.64/MWh. However, this is offset by a consumer surplus loss of $2,135 million per year during peak hours when the consumer price rises from $100/MWh to $150/MWh. Therefore, real-time pricing is attractive to consumers with a demand profile that reflects significant off-peak consumption.

In contrast, demand subscription benefits consumers during peak hours when consumers can hedge price volatility with the flat retail rate and "sell back" unused amounts as demand reduction through a demand-response program when the wholesale market price is high. The consumer surplus is increased by $134 million per year during the peak period. In addition, the consumer can buy additional electricity during off-peak hours when the wholesale market price is low. During the off-peak hours, the consumer surplus is increased by $163 million.

Table 4

Comparing Alternative Demand Reduction Payment Approaches

Case description	**Case 3** **Traditional**	**Case 4** **Second-Best**	**Case 5** **Contract-Based**
Peak hour (6 hours)			
Demand (MW)	17,419	19,000	19,000
Wholesale price ($/MWh)	$134.19	$150.00	$150.00
Retail price ($/MWh)	$94.86	$96.23	$98.18
%Δ in peak demand from Case 1	-12.91%	-5.00%	-5.00%
Δ CS from Case 1 ($million/year)	$620	$229	$134
Δ PS from Case 1 ($million/year)	($1,058)	($427)	($427)
Off-Peak hour (18 hours)			
Demand (MW)	9,501	10,038	10,364
Wholesale price ($/MWh)	$55.01	$60.38	$63.64
Retail price ($/MWh)	$94.86	$96.23	$98.18
%Δ in off-peak demand from Case 1	-4.99%	0.38%	3.64%
Δ CS from Case 1 ($million/year)	$438	$248	$163
Δ PS from Case 1 ($million/year)	-$320	$25	$243
All Hours			
Energy consumption (GWh/Year)	100,571	107,558	109,699
Δ CS from Case 1 ($million/year)	$1,058	$477	$297
Δ PS from Case 1 ($million/year)	($1,377)	($402)	($184)
Δ SS from Case 1 ($million/year)	**($320)**	**$75**	**$113**

Tables 3 and 4 summarize the economic evaluation of the five cases. Not surprisingly, pure price-responsive demand (Case 2) is more efficient than uniform pricing (Case 1), with higher economic surplus, including higher consumer surplus. Price responsiveness in only the high-priced periods (Case 4) also achieves efficiencies, but only 66% as much because it does not avoid under-consumption during low-priced periods. The more surprising and important insight is that the traditional approach (Case 3) is the least-efficient approach, substantially less efficient even than the status quo. The social welfare loss from inducing under-consumption in every period more than offsets the gain from avoiding excessive consumption during peak periods. Some observers will note that consumer surplus appears to increase to the detriment of suppliers, but this is likely to be transient in a competitive market and will disappear as suppliers adjust their investment/retirement decisions, as discussed next in the long-term analysis.

4.2 Long-Run Analysis

It is important to note that the analysis presented above generally represents a short-term equilibrium of the electricity markets. The long term differs from the short term in that suppliers are able to adjust their investment/retirement decisions and their offers into the capacity market such that the price effects of price-responsive demand become much smaller. These long-term effects can be represented in the same economic framework as that used in the short-term analysis by assuming a much higher long-term elasticity of supply reflecting producers' ability to adjust their capacity in the long term. The long-term analysis demonstrates that the relative ranking of alternative approaches is likely to be the same in the long term as in the short term.

The following example shows how incentives can affect economic efficiency in the long term. On the demand side, we assume the same demand functions during peak load and off-peak periods. On the supply side, instead of a linear short-term supply function, we assume that there is a single type of generator with an operating cost of $60/MWh and an amortized capital cost of $100/MWh, which can be recovered through scarcity pricing in the wholesale energy market during the peak period.

Table 5 summarizes the results for Cases 1 through 5 under the long-term assumptions. In Case 1, all demand is insensitive to the real-time wholesale market price. The wholesale market price would be $60/MWh off-peak and $160/MWh peak, the uniform fixed retail rate $100/MWh, the peak load 20,000 MW, and the off-peak load 10,000 MW. In Case 2, all demand is responsive to the wholesale price. When all demand is responsive to the wholesale price, the market demand will drop during the peak period from the initial level by 6%, to 18,800 MW, and rise above 10,000 MW during the off-peak period by 4%, to 10,400 MW. Price-responsive demand improves the economic efficiency and increases the consumers' surplus by $131 million per year.

Next, consider Case 3, in which all demand is price-responsive with extra incentive payments. The extra incentive of the demand reduction payment plus bill savings induces 3,425 MW of demand reduction, which represents a 17% reduction from Case 1. If this is a surprise to the suppliers, there could be transient effects. The demand reduction would create an excess capacity that could drive the wholesale price to the $60/MWh operating cost in the short term. This price drop would cause large transfer benefits during the peak period in the form of bill savings for all customers. The situation is unsustainable, however, because the generation cannot remain commercially viable at such a price. The situation of excess capacity can be alleviated over time by demand growth, plant retirements, unit mothballing, and delays of new generation. In the long run, the price must return to the equilibrium level, as the basic premise of market competition, and the transient gains for the consumers will disappear.

Assuming that the generation capacity adjusts to a new equilibrium level to meet the peak demand, peak prices will return to $160/MWh (with off-peak prices still at $60/MWh) to recover the long-run costs. The retail rate, which reflects the average cost of service (including funding the extra incentive payments for load reductions), rises to $111/MWh. With the extra incentive, the demand responds to the sum of wholesale and retail prices, which equals $271/MWh during the peak period. As a result, the peak demand for this case is reduced to 16,575 MWh, or 17% lower than the level in the uniform pricing case, and 12% lower than the optimal level in the real-time pricing case. Even more strikingly, the off-peak demand is 7% lower than the uniform pricing case and 11% lower than the optimal level in the real-time pricing case, when it should be more efficient to increase demand during the off-peak.

As shown in Table 5, the traditional approach (Case 3) would reduce both the social surplus *and the consumer surplus* by an amount more than four times (Case 2) larger than the economic benefit from real-time pricing. By contrast, if demand is responsive to real-time market prices that exceed the retail rate, but *without* the extra incentive of avoiding retail payments while receiving the LMP (Case 4), social surplus and consumer surplus increase by 60% as much as when demand is always responsive (Case 2). The first-best approach (Case 5) achieves 100% efficiency while attracting the participation of peak demand to become price-responsive.

Table 5 Long-Term Analysis Results

	Case 1 Price Insensitive	Case 2 Price Responsive	Case 3 Traditional Approach	Case 4 Second-Best	Case 5 Contract-Based
Average consumer payment ($/MWh)	100	98	111	100	100
Peak demand (MW)	20,000	18,800	16,575	18,800	18,800
Δ consumer surplus from Case 1 ($million/year)	-	131	(546)	79	131
Δ producer surplus from Case 1 ($million/year)[40]	-	0	0	0	0
Δ Social surplus from Case 1 ($million/year)	-	131	(546)	79	131

[40] The social surplus equals consumer's surplus because the producer surplus is zero because of free entry.

Again, one of the most interesting insights is that the traditional approach (Case 3) could be substantially less efficient than status quo pricing (Case 1). The welfare loss from inducing under-consumption in every period more than offsets the welfare gain from avoiding excessive consumption during peak periods. This result has not taken into account the full effects of moral hazard and adverse selection, as well as activities that exploit free arbitrage opportunities. In Case 3, although consumer surplus may increase to the detriment of suppliers in the short run, this is likely to be transient in a competitive market and will diminish in the long term as suppliers adjust their investment/retirement decisions, as discussed above. Case 4 represents a second-best solution when the retail rate structure remains to be based on fixed uniform pricing for a full delivery service. In Case 5, demand subscription restores efficient incentives for demand-response programs.

5. Conclusion

This paper has reviewed the implications of different approaches to price-responsive demand in electricity markets. Given the recent policy focus on demand response, the paper provides two key insights: First, given administratively-determined customer baselines without financial obligation, introducing demand response can actually reduce economic efficiency relative to the benchmark of uniform retail rates. The efficiency losses arise from the distorted incentives associated with double payment. In addition, demand response could create incentives for gaming strategies that result in illusory demand response. Second, implementing demand response using the demand subscription approach with a contract-based customer baseline generally can achieve similar efficiency results to real-time pricing but with different allocative impacts depending on the range of tariffs available to consumers.

References

Borenstein, S. and Holland, S. (2005) “On the Efficiency of Competitive Electricity Markets with Time-Invariant Retail Prices” *RAND Journal of Economics*, Vol. 36, pp. 469–493.

Borenstein, S., Jaske, M. and Rosenfeld, A (2002) “Dynamic Pricing, Advanced Metering and Demand Response in Electricity Markets”, UCEI, Berkeley, CA.

Brennan, T. J. (2004) “Market Failure in Real-time Metering” *Journal of Regulatory Economics*; 26:2 pp 119-139

Bushnell, Hobbs, and Wolak (2009), “When it comes to Demand Response, is FERC its Own Worst Enemy?,” *The Electricity Journal*, Volume 22, Issue 8, October, pages 9-18

Chao, H. (1983) “Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty” *Bell Journal of Economics*, 14(1), 179-190.

Chao, H. (1989) “Product Differentiation in the Electric Power Industry,” *Proceedings of the 1989 IEEE Conference on Systems, Man, and Cybernetics*, November. *Service Opportunities for Electric Utilities*, Edited by S. Oren and S. Smith (eds.). Kluwer Academic Publishers, Boston, (1993).

Chao, H. (2009), “An Economic Framework of Demand Response in Restructured Electricity Markets” ISO New England, Holyoke, MA. See http://www.hks.harvard.edu/hepg/Papers/2009/Demand%20Response%20in%20Restructured%20Markets%2002-08-09.pdf

Chao, H. (2010) “Price Responsive Demand Management for a Smart Grid World,” *Electricity Journal* (January-February)

Chao, H., Oren, S. and Wilson, R. (2008) “Reevaluation of Vertical Integration and Unbundling in Restructured Electricity Markets” in *Competitive Electricity Markets*, edited by P. Sioshansi, Elsevier Ltd.

Chao, H., Oren, S. and Wilson, R. (2005) “Restructured Electricity Markets: A Risk Management Approach” Electric Power Research Institute, Palo Alto, CA.

Chao, H., Oren, S., Smith, S., and Wilson, R. (1986) “Multilevel Demand Subscription Pricing for Electric Power” *Energy Economics* 8: 199-217.

Chao, H. and Wilson R. (1987) “Priority Service: Pricing, Investment, and Market Organization.” *American Economic Review*, Vol. 77, pp. 899–916.

Cicchetti, C. and W. W. Hogan (1989) “Including Unbundled Demand-side Options in Electric Utility Bidding Programs”, Public Utility Fortnightly, June 8.

Crew, M. and Kleindorfer, P. (1978) “Reliability and Public Utility Pricing”, American Economic Review, Vol. 68, pp. 31-40.

Crew, M. and Kleindorfer, P. (1981) “Regulation and Diverse Technology in the Peak Load Problem,” *Southern Economic Journal*, October

Crew, M., Fernando C. and Kleindorfer, P. (1995) “The Theory of Peak-Load Pricing: A Survey.” *Journal of Regulatory Economics*, 8 3, 215-248.

EPRI (1986) “Priority Service: Unbundling the Quality Attributes of Electric Power”, EA-4851, Project 2440-2, November.

EPRI (2004), *Electricity Sector Framework for the Future,* Electric Power Research Institute Report, Palo Alto, CA.

EPRI (2005), *Electricity Market Transformation: A Risk Management Approach*, Electric Power Research Institute Report, Palo Alto, CA.

Faruqui, F., H. Chao, V. Niemeyer, J. Platt, K. Stalkopf (2001) “*Economics of California’s Power Crisis,*” The Energy Journal.

Faruqui, A., Hledik, R., Newell, S. and Pfeifenberger, H. (2007) “The Power of five percent”, *Electricity Journal*, Vol. 20, Issue 8, pp. 68-77.

Faruqui, A. and Sergici, S. (2009) “Household Response to Dynamic Pricing of Electricity: A Survey of the Experimental Evidence”, The Brattle Group Report. See http://www.hks.harvard.edu/hepg/Papers/2009/The%20Power%20of%20Experimentation%20_01-11-09_.pdf

FERC (2006) “Demand Response and Advanced Metering”, Federal Energy Regulatory Commission Staff Report, Docket AD06-2-000, Washington D.C.

FERC (2008) Wholesale Competition in Regions with Organized Electric Markets, Order No. 719, 73 Fed. Reg. 64,100 (Oct. 28, 2008), FERC Stats. & Regs. P 31,281 (2008) (Order No. 719 or Final Rule).

FERC (2009) “A National Assessment of Demand Response Potential” Federal Energy Regulatory Commission, Staff Report; prepared by The Brattle Group Freeman, Sullivan & Co Global Energy Partners, LLC, June.

Hirst, Eric & Brendan Kirby (2001), *Retail-Load Participation in Competitive Wholesale Electricity Markets*, Jan. 2001, at http://webapp.psc.state.md.us/Intranet/Reports/SU_CM_appendix_A.pdf (prepared for the Edison Electric Institute and Project for Sustainable FERC Energy Policy).

Hogan, W. W. (2001) Hearing on “*FERC: Regulators in a Deregulated Electricity Market*,” before the United States House of Representatives Subcommittee on Energy Policy, Natural Resources and Regulatory Affairs, August 2.

Hogan, W. W. (2009) “Providing Incentives for Efficient Demand Response”, Prepared for Electric Power Supply Association, Comments on PJM Demand Response Proposals, Federal Energy Regulatory Commission, Docket No. EL09-68-000

Hogan, W. W. (2010) “Implications for Consumers of the NOPR’s Prposal to Pay the LMP for All Demand REsponse”, Prepared for Electric Power Supply Association, Comments on Demand Response Compensation in Organized Wholesale Energy Markets, Notice of Proposed Rulemaking, Federal Energy Regulatory Commission, Docket No. RM10-17-000.

ISO-NE (2009) “Status Report on the Future of Price-Responsive Demand Programs Administered by ISO New England Inc.” February 13. See http://www.iso-ne.com/committees/comm_wkgrps/mrkts_comm/mrkts/mtrls/2009/feb202009/a2_iso_status_report_prd_draft_version_1_02_13_09.pdf

Joskow, P. (2001) “*California’s Electricity Crisis,*” December 2001, Oxford Review of Economic Policy, Volume 17.

Joskow, P. and Tirole, J. (2006) “Retail Electricity Competition” *RAND Journal of Economics*, Vol. 37 pp. 799–815.

Joskow, P. and Tirole, J. (2007) “Reliability and Competitive Electricity Markets” *RAND Journal of Economics*, Vol. 38, No. 1, pp. 60–84.

Lovins, A. B. (1985) “Saving Gigabucks with Negawatts,” *Public Utilities Fortnightly*, Vol. 115, No. 6, Mar. 21, p. 24

Panzar, J. C. and Sibley, D. S. (1978) “Public Utility Pricing Under Risk: The Case of Self-Rationing”, American Economic Review, Vol. 68, No. 5, pp. 888-895.

Rochlin, C. (2009) “The Alchemy of Demand Response: Turning Demand into Supply” Center for Research in Regulated Industries, 22nd Annual Western Conference, Monterey, California, June 19.

Ruff, L. (2002) “Economic Principles of Demand Response in Electricity”, Edison Electric Institute, Washington D. C.

Schweppe, F. C., Tabors, R. D. and Kirtley, J. (1982) “Homeostatic Control for Electric Power Usage” IEEE Spectrum, July, pp 44-48.

Sweeney, J. (2002) *The California Electricity Crisis*, Hoover Institution Press.

Wellinghoff, J. and Morenoff, D. (2007) “Recognizing the Importance of Demand Response: the Second Half of the Wholesale Electric Market Equation”, *Energy Law Journal*, Vol. 28, No. 2 389-419.

Wilson, R. (1989) “Efficient and Competitive Rationing”, *Econometrica* 57: 1-40.

One of the current problems in restructured electricity markets is the highly inelastic demand for electricity among all but the largest consumers.Â Demand-side resources that are available for deployment when prices are high because of a low generation reserve margin could serve as a shock absorber for end-use customers in the face of the time lag of building new power plants. The thinner the generation margin, the higher market prices would be, which would provide more demand response.Â Load management or â€œpeak shavingâ€ programs, which focus on reducing electricity use during summer afternoons, are more effective when prices are predictably high for enough hours to justify an investment in peak-shaving technologies or processes.