

February 8, 2009

An Economic Framework of Demand Response in Restructured Electricity Markets

Hung-po Chao¹

ABSTRACT

This paper provides a unified economic framework for assessing the effectiveness of price-responsive demand in promoting efficient electricity markets. The framework recognizes the interactions between a competitive wholesale market and regulated uniform retail rates within a hybrid market structure. The framework features the basic principles of marginal cost pricing in the wholesale market and average cost pricing in the retail market while applying the advanced theory of priority service to shortage/scarcity pricing. The analysis addresses the institutional and technical barriers that create an asymmetric information structure between consumers and producers inhibiting efficient price responsive demand in electricity markets. The analysis indicates that while demand response incentives could potentially alleviate the inefficiencies before these barriers are removed, the use of Customer Baseline is susceptible to gaming problems, and it could also create an excessive incentive inadvertently causing inefficient price formation. If not corrected, this could result in inefficiencies that would outweigh the benefits of demand reduction during the peak period and increase the average cost to the final consumers.

1. INTRODUCTION

The importance of demand response for a successful reform of electricity markets has long been recognized. (Joskow, 1983) During the past two decades, however an unbundled market

¹ Director, Market Strategy and Analysis, Market Monitoring Unit, ISO New England, One Sullivan Road, Holyoke, MA 06029. I am very grateful to the generous comments from and the productive exchanges with many of my colleagues at ISO New England, NEPOOL Market Participants, the staff of Public Utility Commissions and other government offices in New England states and individuals in academic institutions, including the participants at a workshop held by MIT Center for Energy and Environmental Research (CEEPR) on November 21, 2008. The current paper is informed by a study that is still in progress. I appreciate Kamen Madjarov and Gail Adams for their capable assistance with the modeling, while I am solely responsible for any remaining errors. The views expressed herein are those of mine and do not necessarily represent the positions of the ISO or the views of others.

structure with a hybrid of competitive wholesale market and regulated retail markets has emerged in the U.S., presenting new challenges for the development of demand response. In a recent paper, Chao, Oren and Wilson (2008) re-evaluated vertical integration and unbundling in restructured electricity markets and concluded that arguments can be mustered for either structure without any definitive conclusion that one or the other extreme is better. Therefore, the current hybrid structure with regulated retail rates and competitive wholesale prices is likely to persist as the industry evolves. With a hybrid market structure, demand response takes on special importance for it serves as a critical link between the wholesale and retail markets. Given the strategic importance of demand participation in electricity markets, Wellinghoff and Morenoff (2007) contend that both the federal and state governments have strong roles for assuming jurisdiction over facilitating the development of demand response. This paper addresses some new challenges in light of the growing prospect of demand response in restructured electricity markets.

The U. S. Department of Energy (DOE) defines demand response in its February 2006 Report to Congress:

“Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized.”

In general, there are a variety of mechanisms that promote price-responsive demand in both wholesale and retail markets. The terms “demand response” and “price-responsive demand” have often been used interchangeably, though there are important distinctions as well. Price-responsive demand is a customer’s ability to alter its electricity demand by reducing or shifting consumption in response to market prices or other market conditions. Programs promoting price-responsive demand in retail markets are predominantly time-based pricing programs, including real-time pricing and other dynamic retail pricing programs, whereas demand response

programs in wholesale market can be designed to promote system reliability and economic efficiency.²

The economic theory of demand response is rooted in the peak load pricing literature. In the early literature, Steiner (1957) and Boiteux (1960) introduced the peak load pricing and investment issue in the presence of price-responsive demand. Subsequently, Brown and Johnson (1969), Carlton, D. (1977) and Crew and Kleindorfer (1976, 1978) and Chao (1983) analyzed the pricing and investment issues in the presence of demand and supply uncertainties. Chao and Wilson (1987) and Wilson (1989) studied priority pricing as an incentive mechanism for efficient contingent allocation of electric energy through auction and forward contracting. More recently, Borenstein and Holland (2005) and Joskow and Tirole (2006, 2007) studied retail competition in restructured electricity markets. Borenstein, Jaske and Rosenberg (2002) and Ruff (2002) reviewed the economic principles of demand response.

This paper presents a unified economic framework on electricity pricing and investment within a hybrid market structure that encompasses competition in the wholesale market and regulation in the retail market. This paper presents an economic framework to advance understanding of the effects of price responsive demand on wholesale and retail markets, for the purpose of exploring concepts to improve program design and evaluating alternative designs. The analysis establishes the socially optimal pricing and investment decisions within a hybrid market structure: The optimal uniform retail price equals to the marginal-demand-weighted expectation of the competitive wholesale prices. Efficient investment is described by the standard free entry condition. The optimal capacity level strikes a balance between the cost of shortage and the cost of excess capacity. To foster efficient rationing during shortage periods, the analysis incorporates the advanced theory of priority service to facilitate an incentive-based demand response mechanism for efficient scarcity pricing. The optimal allocation within a

² Demand-response programs that focus on system reliability generally provide the power system operator the ability to call customers to take their electrical load off of the bulk power system when the system is deficient in capacity or operating reserves. Since some customers are able to take electrical load off the grid very quickly and are often willing to do so at prices less than the cost of building new generating capacity, such demand response programs can be very cost-effective sources of capacity or reserves to wholesale markets. Demand response programs that focus on price-responsive demand place a greater emphasis on using wholesale energy market price signals to improve the economic efficiency of energy consumption.

hybrid market structure can be implemented through a competitive wholesale market with a two-part retail tariff.

The analysis recognizes the asymmetric information structure between consumers and producers with respect to the real-time prices. This reflects the institutional and technical realities associated with the default uniform retail rate and inadequate advanced metering infrastructure that enable demand management in response to real-time prices. In other words, a consumer's demand response ability is constrained. These constraints cause two market failures: First, since consumers cannot respond to real-time prices, the real-time demand is inelastic. This creates a moral hazard problem. As a consequence, the wholesale market is susceptible to potential market power by producers when the system capacity is tight. Second, since the local utility or load serving entity cannot bill retail customers for their incremental costs of services with different real-time demand profiles, this creates an adverse selection problem. As a result, during a peak-load period, customers over consume electricity with values that are often below the marginal cost, and during an off-peak period, customers under consume electricity with values that exceed the marginal cost. Given the reality of an imperfect hybrid market structure, there is potential for efficiency improvement through demand response incentives. However, the analysis suggests that to enhance efficiency, both the existing and new approaches need to be evaluated in recognition of these constraints.

The remaining sections of this paper are organized as follows. Section 2 presents the basic model of a hybrid market structure with competitive wholesale market and optimal retail regulation under uniform pricing. Section 3 presents barriers to price responsive demand. Section 4 addresses the demand response incentives issues associated with the use of customer baseline. Section 5 concludes with a summary of the main observations.

2. The Basic Model

This section presents the basic model of a hybrid market structure consisting of a competitive wholesale real-time energy market and a regulated uniform fixed retail rate with

price insensitive demand. The basic model builds on the previous work of Chao (1983), Chao and Wilson (1987), Borenstein and Holland (2005), and Joskow and Tirole (2006) among others.

Ideally, the wholesale and retail markets are efficient when the marginal cost of energy equals the real-time energy market price in a competitive wholesale energy market, and retail consumers can see these prices and may adjust their real-time consumption in response to these prices – i.e., their demand is price responsive.³ However, we begin with the more realistic assumption that electric utilities offer retail consumers at a uniform retail rate that does not vary with time, while the wholesale market produces competitive energy prices that varies from hour to hour. Uniform pricing was expedient when the metering technology was primitive. The traditional electric meter measures the total kWh usage that is only adequate for billing customers under a flat rate. With traditional meters, consumers can not be charged real-time prices that reflect the marginal costs of production. Therefore, the demand is inelastic or insensitive to real-time prices that reflect marginal costs. The uniform retail rate represents an average cost per unit of electricity consumed, including components for the recovery of investment costs. Therefore, retail electricity demand is responsive only to the flat rate that reflects the cost of service for all rate payers averaged over a rate period but does not reflect the marginal cost of electricity production in real time.

On the other hand, the wholesale market is grounded in the technical reality that electricity is a time-differentiated product that cannot be stored economically at the present time. Generation units are dispatched in real-time to meet constantly fluctuating demand. To minimize the cost of operating an interconnected power system, generation units are generally dispatched in merit order, starting with those with the lowest operating costs, and progressively dispatching more expensive generators as the demand increases in real time. As a result, the marginal cost of energy fluctuates from hour to hour in real time.

³ The real-time wholesale energy market price is also known as “the market clearing price of energy,” “the spot price of energy,” or “the Locational Marginal Price” in the industry

Model Assumptions

In this paper, we focus on the energy market and ignore the transmission security issue. We assume that the transmission capacity is abundant and reliable; rationing through load shedding is adequate to avert the possibility of cascading failures and system collapse.

We consider the stochastic nature of demand and supply. Let $(\Theta, \mathfrak{S}, \mathcal{F})$ denote a complete probability space, \mathfrak{S} is a σ -field on Θ , and \mathcal{F} a probability measure on \mathfrak{S} . The states of nature are represented by a vector $\theta \in \Theta$. The time-related variation and the stochastic variation of electricity demand can be treated in the same manner. Let $E_\theta[\cdot]$ denote expectations with respect to the probability measure \mathcal{F} .

A. Heterogeneous demand

We assume that retail consumers are heterogeneous with different tastes and prefer different demand profiles. The consumer population is denoted by L . The money-metric utility function of energy consumption for a consumer of type $l \in L$ is denoted by $U_{l\theta}(q_{l\theta})$. We assume that $U_{l\theta}(\cdot)$ is an increasing concave function with $U_{l\theta}(0) = 0$. The consumer's behavior is described by the condition that marginal utility, $U'_l(q_{l\theta})$, equals the price of electricity when the individual demand level is $D_{l\theta}(p)$, or $U'_l(D_{l\theta}(p)) = p$. The market demand is the sum of individual demand functions, $D_\theta(P) = \sum_l D_{l\theta}(P)$. For any price P , the gross value function can be written as $U_\theta(D_\theta(P)) = \sum_l U_{l\theta}(D_{l\theta}(P))$, or equivalently,

$$U_\theta(D_\theta(P)) = \max_{q_{l\theta}} \{\sum_l U_{l\theta}(q_{l\theta}) \mid D_\theta(P) = \sum_l q_{l\theta}\}.$$

Since consumer preferences can be differentiated in many dimensions, there is no workable general theory for consumer preferences for differentiated products. In this paper, I focus on the time differentiation and consider a preference structure that captures some essential ingredients of price responsive demand. Specifically, I assume that the individual consumer's utility function takes on the following structure:

$$U_{l\theta}(q_{l\theta}) = \varphi_{l\theta} u_l(q_{l\theta}/\varphi_{l\theta}). \quad (1)$$

This implies that $U'_l(q_{l\theta}) = u'_l(q_{l\theta}/\varphi_{l\theta})$. Letting the marginal utility equal to the electricity price, p , we obtain $D_{l\theta}(p) = \varphi_{l\theta} D_l(p)$, where $u'_l(D_l(p)) = p$. Therefore, each consumer's

demand can be expressed as the product of a state-contingent component $\varphi_{l\theta}$ and a price-sensitive component, $D_l(p)$. The separable structure of the utility function allows us to identify consumers' preferences based on the demand profile, $\varphi_{l\theta}$. Theoretically, this preference structure simplifies the presentation of general results and sharpens the insights. Practically, it means that a consumer's utility function can be estimated from the observable data.

Shortage pricing

The stochastic nature of demand and supply gives rise to the possibility of shortage events and the necessity for non-price rationing. Let $\rho_\theta \in [0,1]$ and $\rho_{l\theta} \in [0,1]$ denote, respectively, the fractions of market demand and of consumer l 's demand that are served in state θ . We assume that $\rho_{l\theta}D_{l\theta}(P)$ is the actual demand for consumer l in state θ after rationing during a shortage event and that the each individual consumer's gross surplus is proportional to the fraction of demand being rationed, $\rho_{l\theta}U_{l\theta}(D_{l\theta}(P))$, implying that rationing generally results in non-optimal energy consumption, or $\rho_{l\theta}U_{l\theta}(D_{l\theta}(P)) \leq U_{l\theta}(\rho_{l\theta}D_{l\theta}(P))$ when $\rho_{l\theta} < 1$. We assume that the value of lost load for consumer l equals the average utility of consumption,⁴

$$VOLL_{l\theta} = \frac{U_{l\theta}(D_{l\theta}(P))}{D_{l\theta}(P)}.$$

Every feasible rationing plan $(\rho_{l\theta})$ satisfies the condition that the actual market demand equals the sum of the actual individual demands, or $\rho_\theta D_\theta(P) = \sum_l \rho_{l\theta} D_{l\theta}(P)$. The realized gross value function is written as, $U_\theta(D_\theta(P), \rho_\theta) = \sum_l \rho_{l\theta} U_{l\theta}(D_{l\theta}(P))$.

We define the value of lost load ($VOLL$) as the marginal utility that is reduced due to rationing,

$$VOLL_\theta = \frac{\partial U_\theta(D_\theta(P), \rho_\theta) / \partial \rho_\theta}{D_\theta(P)}$$

For illustration, we consider two special cases of rationing: 1) random rationing and 2) optimal rationing.

⁴ The linear formulation is not restrictive. It can be easily generalized so that a consumer's utility function depends on rationing in a nonlinear fashion, say, $U_{l\theta}(q_{l\theta}, \rho_{l\theta})$. There could be multiple demand blocks, within each of which the value of lost load is defined linearly.

1) Random rationing

For random rationing, we represent it through two requirements: 1) the each individual consumer's gross surplus takes the linear form of being proportional to the fraction of demand rationing, $U_{l\theta}(D_{l\theta}(P), \rho_{l\theta}) = \rho_{l\theta} U_{l\theta}(D_{l\theta}(P))$, which implies that consumers do not manage demand reductions efficiently during shortage events, and 2) all consumer demands are rationed indiscriminately with an equal probability, $\rho_{l\theta} \equiv \rho_{\theta}$. For this case, we obtain, $VOLL_{\theta} = U_{\theta}(D_{\theta}(P))/D_{\theta}(P)$.

2) Optimal rationing

In this case, we assume that the optimal rationing plan ($\rho_{l\theta}^*$) maximizes the gross surplus:

$$U_{\theta}(D_{\theta}(P), \rho_{\theta}) = \max_{\rho_{l\theta}} \left\{ \sum_l \rho_{l\theta} U_{l\theta}(D_{l\theta}(P)) \mid \rho_{\theta} D_{\theta}(P) = \sum_l \rho_{l\theta} D_{l\theta}(P) \right\}.$$

The optimal rationing plan is to allocate the available supply during a shortage event to consumers with the highest value of lost load. The optimal rationing plan can be characterized as follows: the available resource supply is allocated to demand in the decreasing order of consumer's value of lost load, $v_l(P) = u_l(D_l(P))/D_l(P)$, until the supply is exhausted.

Therefore, we have,

$$\rho_{l\theta}^* = \begin{cases} 1, & \text{if } v_l(P) > v_{l^*}(P) \\ \rho_{l^*\theta} \in (0,1), & \text{if } v_l(P) = v_{l^*}(P). \\ 0, & \text{if } v_l(P) < v_{l^*}(P) \end{cases}$$

With optimal rationing, we obtain

$$VOLL_{\theta} = \frac{\partial U_{\theta}(D_{\theta}(P), \rho_{\theta}) / \partial \rho_{\theta}}{D_{\theta}(P)} = v_{l^*}(P) \quad (2)$$

During a shortage event, the marginal rationing cost provides the basis for efficient shortage pricing, if information on the value of lost load, $v_l(P)$, can be solicited from individual consumers.

Note that the value of lost load function $v_l(P)$ is independent of the state of nature. Therefore, in principle, it is possible to obtain its value in advance to form a state-contingent forward contract for rationing and pricing during shortage periods. We will return to this later in the paper.

B. Supply technology

On the supply side, we assume that there is a continuum of investment opportunities with different amortized capital cost of unit capacity investment, k , and operating cost c .⁵ The investment opportunities are represented by the unit capacity cost, k , including the cost of capital, as a function of the operating cost, $k(c)$ for $c \in [\underline{c}, \bar{c}]$. We assume that this function is decreasing in c . If demand were certain and constant, the lowest cost technology, namely the base load unit, would be the only supply option that gets built to meet the load. But when load varies stochastically, and output is not storable, idled base-load capacity is expensive. Some peaking units may be off-line most of the time, running only for very low-probability contingencies that cover as few as 10 to 50 hours each year. The efficient result is to invest in a balanced portfolio of technologies with varying capacity and operating costs that minimize the overall production cost.

Let $s(c)$ denote the capacity level of units with an operating cost c . Let $\alpha_\theta(c)$ be the availability factor of unit c in the state θ . We suppose that all available units will be run in economic merit order from the lowest operating cost to higher costs. The supply function can be obtained as the cumulative output from available units with operating costs below the price, p ,

$$S_\theta(p) = \int_0^p \alpha_\theta(c) s(c) dc.$$

The total capacity investment cost is

$$I = \int_0^1 k(c) s(c) dc$$

The total operating cost is

⁵ Chao (1983) employed a discrete model to represent investment opportunities. With no loss of generality, we adopt the more elegant continuous model by Joskow and Tirole (2007). The two different models produce fundamentally the same results.

$$C_\theta(S_\theta(p)) = \int_0^p c \alpha_\theta(c) s(c) dc.$$

C. Social welfare function

In the basic model, we adopt the economic surplus as the standard measure of social welfare function.⁶ Given the informational constraint, a Ramsey planner chooses a uniform fixed retail price P , a set of wholesale prices (P_θ), a capacity plan $s(c)$ and a rationing plan ($\rho_{l\theta}$) that maximize the expected total economic surplus:

$$\text{Max } E[U_\theta(D_\theta(P), \rho_\theta) - C_\theta(S_\theta(P_\theta))] - \int_0^{\bar{c}} k(c) s(c) dc$$

$$S_\theta(P_\theta) \geq \rho_\theta D_\theta(P)$$

Assumption: $E[(P - P_\theta)D'_\theta(P)]$ is decreasing in P .

This technical assumption ensures the sufficiency of the second order optimality condition. The first-order optimality condition for a hybrid market structure can be characterized as efficient pricing and efficient investment conditions.

First, efficient pricing and allocation dictates a) the determination of wholesale market price in relation to marginal costs and value of lost load (VOLL) and b) the determination of the uniform fixed retail rate in relation to the wholesale market prices under the condition the market clearing condition of demand and supply.

$$S_\theta(P_\theta) = \rho_\theta D_\theta(\hat{P}),$$

where

$$\rho_\theta = \text{Min}\{S_\theta(\bar{c})/D_\theta(\hat{P}), 1\}.$$

⁶ Economic 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.

The cleared market demand, $D_\theta(\hat{P})$, is the point on the demand function where the price equals the uniform retail rate, \hat{P} . The wholesale price, P_θ , is set at the level where the supply equals the cleared market demand. Rationing is activated, $\rho_\theta < 1$, when the cleared market demand exceeds the total available capacity. The efficient pricing results are summarized below:

Result 2.1: The competitive wholesale market price equals 1) the marginal cost of production, during a normal period ($\rho_\theta = 1$) when $S_\theta(\bar{c}) \geq D_\theta(\hat{P})$, or 2) the marginal value of lost load, during a shortage period ($\rho_\theta < 1$) when $S_\theta(\bar{c}) < D_\theta(\hat{P})$. That is,

$$P_\theta = \begin{cases} C'_\theta(S_\theta(P_\theta)), & \text{if } \rho_\theta = 1 \\ VOLL_\theta, & \text{if } \rho_\theta < 1 \end{cases}$$

Result 2.2: The optimal uniform retail price equals to the expected marginal-demand-weighted average of competitive wholesale market prices, $\hat{P} = E[P_\theta \rho_\theta D'_\theta(\hat{P})] / E[\rho_\theta D'_\theta(\hat{P})]$.

Second, efficient investment is characterized by the standard zero-profit free entry condition for new investment:

Result 2.3 The efficient investment plan is characterized by the zero-profit free-entry condition.

$$k(c) = E[\alpha_\theta(c)(P_\theta - c)^+], \quad \text{if } s(c) > 0.$$

It is straightforward to show that $k'(c) < 0$, and this suggests that efficient technology frontier is downward sloping. Applying the efficient investment condition to a peaker, a unit with the highest operating cost \bar{c} and $s(\bar{c}) > 0$, we obtain the condition for the optimal capacity requirement,

$$LOLP = \Pr\{\rho_\theta < 1\} = \Pr\{S_\theta(\bar{c}) < D_\theta(\hat{P})\} = \frac{k(\bar{c})}{VOLL - \bar{c}}$$

$$\text{where } VOLL = E\{VOLL_\theta | \rho_\theta < 1\}.$$

This is summarized in the following result.

Result 2.4: The optimal capacity level is attained when the lost of load probability equals $LOLP = k(\bar{c}) / (VOLL - \bar{c})$.

Example:

Suppose that for a peak load unit,

$$k(\bar{c}) = \$90/\text{kW-year}$$

$$\bar{c} = \$250/\text{MWh}$$

$$\text{VOLL} = \$4000/\text{MWh}$$

The optimal capacity level is achieved when

$$\text{LOLP} = \frac{90/8.76}{4000 - 250} = 0.00274 = 1 \text{ day per year}$$

Result 2.3 suggests that in a competitive wholesale market, a generator should be able to earn adequate returns to recover the costs of capacity investment. However, as we will show in the following, the revenue adequacy condition does not always hold in the retail market, because it is uncertain whether the retail revenue under the optimal uniform retail price is higher or lower than the amount of revenue needed to recover the procurement costs. The only uniform retail price that is revenue neutral in the sense that it will generate the right amount of revenue to recover the procurement cost from the wholesale market is the demand-weighted expectation of competitive wholesale prices.

Result 2.5: The demand-weighted expectation of competitive wholesale market prices is the only uniform retail price that will yield adequate net revenue to recover the operating and capacity investment costs,

$$\bar{P} = \frac{E[\rho_{\theta} P_{\theta} D_{\theta}(\bar{P})]}{E[\rho_{\theta} D_{\theta}(\bar{P})]}.$$

Since \bar{P} may be different from, \hat{P} , the optimal retail rate does not assure adequate revenue to recover the investment costs. When marginal demand is independent of the state of nature, the efficient uniform fixed price flat rate will run deficit. Only in a special circumstance when consumers are homogenous with the same demand profile, φ_{θ} , the marginal demand is perfectly

correlated with the total demand, $D_\theta(p) = \varphi_\theta D(p)$. Borenstein and Holland (2005) show that under Bertrand competition in the retail market with uniform fixed retail prices, the equilibrium retail price equals to the demand-weighted average wholesale price, \bar{P} . Since the resulting allocation is less efficient than the second-best optimal retail price, \hat{P} , Borenstein and Holland conclude that there is market failure in retail competition. Joskow and Tirole (2007) refuted the BH conclusion with the suggestion that the retailer could compete by offering customers two-part tariff.

Economic theory recommends two general approaches to collect fixed revenue from regulated retail rates with minimum inefficiency. One approach is Ramsey pricing or second-best pricing, which suggests that to minimize the overall inefficiencies in customers' choices under a revenue constraint, the prices should deviate from incremental costs in such a way that customers with more inelastic demands pay larger shares of the deficit. The revenue neutral price in BH can be interpreted as Ramsey pricing, because the revenue neutral price can be obtained as a modified optimal price with the marginal demand weights scaled by the inverse of demand elasticity. That is, $E[p_\theta D_\theta(P)]/E[D_\theta(P)] = E[p_\theta D'_\theta(P)/\varepsilon_\theta]/E[D'_\theta(P)/\varepsilon_\theta]$, where $\varepsilon_\theta = D'_\theta(P)P/D_\theta(P)$.

The alternative approach uses a two-part tariff, which is generally considered to be more efficient than single-part linear tariff, like Ramsey pricing. (Brown and Sibley, 1986; Wilson, 1993; Joskow and Tirole, 2006) In a two-part tariff, for example, each customer pays a “demand charge” that depends on the peak load, and then, for each kW within that peak load, an energy charge based on the amount of energy, measured in kWh, that is used during the year. In effect, this scheme charges the customer for its load profile over a billing period. The above observations are summarized in the following result.

Result 2.6 The optimal allocation and investment plan in a hybrid market structure is attained via a competitive wholesale market and an optimal two-part retail tariff.

Scarcity pricing⁷

In the remainder of this section, we consider rationing during a shortage period as a form of reliability-based demand response program. We consider priority service mechanism, a menu of forward contingent contracts which implement shortage pricing for efficient rationing during a shortage period. Priority rationing is uniformly more efficient than random rationing, and there should be no losers by switching from random rationing to priority rationing.⁸ To implement priority rationing, it is essential that the customers are willing to reveal their value of lost load per unit of demand, so that demand reduction can be ranked in an efficient manner that minimizes the total value of lost load. A priority service mechanism is built on the diversity among customers' relative preferences: some want lower prices and are willing to be curtailed more frequently than others whereas others are willing to pay a premium for a more reliable service. Each consumer is asked to report a value of lost load, $v_l(\hat{P})$, in advance, for the purpose of establishing a priority order in case of shortage. During a shortage event, the demand is curtailed according to an increasing order of $v_l(\hat{P})$, with a lower value demand cut off before a higher value demand, until the demand is reduced to a level that can be met with the available capacity. During a shortage period, the real-time price equals the marginal value of lost load for the demand not curtailed, $P_\theta = VOLL_\theta = v_{l^*}(\hat{P})$, where l^* represents the type of the marginal consumer. The curtailed demand will be compensated at a price that equals to $VOLL_\theta - \hat{P}$. The following result shows that this design is incentive compatible and efficient.

Result 2.7 The priority service mechanism implements efficient rationing and scarcity pricing.

Proof:

Since the rationing plan based on the ranking order of value of lost load is efficient, we only need to show that the mechanism is incentive compatible. Suppose that the real-time price is $P_\theta = VOLL_\theta$ during a shortage event. The consumer's surplus is given by

⁷ In general, the terms shortage pricing and scarcity pricing are used interchangeably. The fine distinction made here is that in shortage pricing, the value of lost load is based on an objective estimate approved by the regulator, whereas in scarcity pricing, the value could be obtained through voluntary demand bids, a process which is approved by the regulator.

⁸ See Chao and Wilson (1987) and Wilson (1989).

$$CS_{l\theta} = \begin{cases} [v_l(\hat{P}) - \hat{P}]D_{l\theta}(\hat{P}), & \text{if } v_l(\hat{P}) \geq P_\theta \\ [P_\theta - \hat{P}]D_{l\theta}(\hat{P}), & \text{if } v_l(\hat{P}) < P_\theta \end{cases}$$

If consumer of type l bids a value that is higher than the true value of lost load, i.e., $v > v_l(\hat{P})$, then the only difference it makes is that when $v > P_\theta > v_l(\hat{P})$, the consumer would not be curtailed. However, the opportunity cost to the consumer is to forego the scarcity rent, $[P_\theta - \hat{P}]D_{l\theta}(\hat{P})$, which is higher than the consumer's surplus, $[v_l(\hat{P}) - \hat{P}]D_{l\theta}(\hat{P})$, by the amount, $[P_\theta - v_l(\hat{P})]D_{l\theta}(\hat{P}) > 0$.

On the other hand, if the consumer under-reports the true value of lost load, i.e., $v < v_l(\hat{P})$, then the only difference it makes is that when $v < P_\theta < v_l(\hat{P})$, the consumer will be curtailed and be compensated at a shortage price which is lower than the true value of lost load. Consequently, the consumer's surplus will be reduced by the amount, $[v_l(\hat{P}) - P_\theta]D_{l\theta}(\hat{P}) > 0$.

Therefore, a consumer's best strategy is to report $v_l(\hat{P})$ truthfully, and the mechanism is incentive compatible.

Q.E.D.

Revenue Inadequacy and Price-cap

Price cap in the wholesale market is generally considered as an interim measure introduced to mitigate market power and moderate price volatility during a shortage period. But like any price intervention mechanism, price cap causes inefficiencies. First, by limiting producers' profits during peak periods, it diminishes the incentive for investment in new generating capacity. Second, by capping scarcity prices during shortage periods, it causes inefficient rationing and distorts the price signals. In the following, we study the use of capacity payments as a compensatory market mechanism to offset the undesirable effects of price cap.

Suppose that consumers report the values of lost load ($VOLL_{l\theta}$) truthfully before the real-time market is conducted, the scarcity price $VOLL_\theta$ can be calculated during a shortage event as discussed earlier. We denote by \check{P} the price cap on real-time prices. With a price cap, \check{P} , the wholesale price during a shortage event is modified as follows:

$$P_\theta = \min\{VOLL_\theta, \check{P}\}, \text{ when } \rho_\theta < 1.$$

As a consequence, the price cap lowers a consumer's cost per unit of demand by the following amount,

$$\pi_l(\hat{P}) = \frac{E \left[\rho_{l\theta} \varphi_{l\theta} (VOLL_\theta - \check{P})^+ | \rho_\theta < 1 \right] Pr\{\rho_\theta < 1\}}{E[\rho_{l\theta} \varphi_{l\theta} | \rho_\theta < 1]}$$

The price cap lowers producers' total revenue per unit of available capacity by an amount that reflects the reduced scarcity rent,

$$\pi(\hat{P}) = \frac{E \left[S_\theta(\bar{c}) (VOLL_\theta - \check{P})^+ | \rho_\theta < 1 \right] Pr\{\rho_\theta < 1\}}{E[S_\theta(\bar{c}) | \rho_\theta < 1]}$$

The total amount of capacity payment for producers equals the total priority service charges for consumers:

$$\pi(\hat{P}) E[S_\theta(\bar{c}) | \rho_\theta < 1] = \sum_l \pi_l(\hat{P}) E[\rho_{l\theta} \varphi_{l\theta} D_l(\hat{P}) | \rho_\theta < 1].$$

The capacity payment $\pi(\hat{P})$ recaptures the scarcity rent reduced by the price cap and thus restores the financial incentives for efficient investment.

The amount $\pi_l(\hat{P})$ is what a risk-neutral insurance underwriter would offer a customer in a supplemental insurance contract to compensate them for the financial risks of shortage pricing during a shortage period without the price cap. Suppose that a menu of priority service options $\{\pi_l(\hat{P})\}$ is offered to all consumers, and that the rationing plan will be implemented according to a ranking order based on consumers' selections of priority payment, a consumer whose value of lost load is $v_l(\hat{P})$ is predicted according to Result 2.7 to select the priority payment, $\pi_l(\hat{P})$. Consequently, consumers will reveal their true value of lost load.

In principle, price cap may affect not only the scarcity prices but also real-time prices in the wholesale market during a normal period. Therefore, an efficient allocation must be predicated upon a sufficiently high price cap. Therefore, we may summarize the result generally as follows.

Result 2.8 If the price cap is sufficiently high so that it does not interfere real-time pricing, efficient allocation can be attained through a supplemental capacity payment and priority payments.

3. Challenges to Price Responsive Demand

Despite the technical advances and political support over the past two decades, the progress of demand response has been modest. (FERC 2006) In this section, we discuss the importance of price responsive demand in a hybrid market structure and enumerate barriers to its development.

Price-responsive demand improves the economic efficiency by discouraging low value energy consumption when real-time wholesale energy prices are high during the peak period and encouraging high value energy consumption when real-time wholesale energy prices are low during the off-peak period. Price-responsive demand could reduce demand volatility and the peak demand. This, in turn, reduces the need to install additional generation and transmission infrastructure to serve the peak. Price responsive demand reduces the likelihood of calling on operating reserves and the likelihood of involuntary load shedding. It also reduces the ability of suppliers to exercise market power improving competition in energy markets and lowering prices in constrained market areas served by few generation companies and tight supply situations. Facing an inelastic demand, producers can profitably increase prices with little or no loss of sales. When the demand for electricity is responsive to changes in the wholesale price, an attempt to withhold or raise prices will cause reduction in sales, thus reducing the profitability of such an action and discouraging the exercise of market power. When the retail price of electricity does not vary with wholesale market prices, however, consumers do not see changes in the (retail) price they face, and thus the demand is inelastic despite the run-up in wholesale prices.

Various recent surveys provide empirical evidence that customers respond to the higher prices by lowering demand. For instance, in a recent survey of seventeen pricing experiments,⁹ Faruqui and Sergici (2008) conclude:

“Across the range of experiments studied, time-of-use pricing tariffs lead to a drop in peak demand that range between 3 and 6 percent and critical-peak pricing tariffs lead to a drop in peak demand of 13 to 20 percent. When accompanied with enabling technologies, the latter set of tariffs leads to a drop in peak demand in the 27 to 44 percent range.”¹⁰

Linking real-time competitive wholesale market energy prices and retail rates would encourage price-responsiveness among retail customers. In reality, however, the vast majority of consumers remain largely on uniform fixed retail rates, forming a barrier to price-responsive demand. The importance to remove the institutional barriers and to engage customers on the demand-side of the electricity market is widely recognized as one of the most important lessons from the California electricity crisis. As Joskow (2002) aptly observed:

“The answer for California now is not to return to the old, costly system of regulated monopolies, but to apply the harsh lessons it has learned from designing a flawed system. Competitive electricity markets will not work if consumers are completely insulated from wholesale market prices.”

In the following, we discuss the two barriers that prevent the full potentials of price-responsive demand from being realized: the default uniform retail rates and the lack of advanced metering infrastructure.

⁹ Faruqui and Sergici (2008), “Household Response to Dynamic Pricing of Electricity: A Survey of Seventeen Pricing Experiments”, The Brattle Group, San Francisco, CA, and Cambridge, MA. *See* [http://www.hks.harvard.edu/hepg/Papers/2009/The%20Power%20of%20Experimentation%2001-11-09 .pdf](http://www.hks.harvard.edu/hepg/Papers/2009/The%20Power%20of%20Experimentation%2001-11-09.pdf)

¹⁰ The 2007 peak load in New England was 26,145 MW. A 20 percent of peak demand amounts to 5,229 MW.

First, a fixed-price uniform retail rate impedes price-responsive demand in wholesale and retail markets, and has adverse economic and institutional ramifications. The inefficiencies caused by fixed-price uniform rates can be enumerated as follows:

- A fixed-price uniform retail rate causes a number of economic inefficiencies because it does not enable different customers to express the individual value they place on electricity, and does not reflect the true cost of serving individual customer demand. Electricity customers have different preferences for quality of service, load profiles and demand curves. For instance, a uniform rate creates a tension between those customers who might prefer lower rates with higher risks of outage and those who prefer more reliable service at higher rates.
- A uniform rate causes volatile and inelastic levels of demand. Because a uniform rate does not reflect directly the incremental costs of services demanded by different customers in response to that rate, electricity may be produced and delivered at costs that exceed the value of electricity to its customers or may be withheld from customers with values that exceed the cost of production. In practical terms, the demand for electricity and wholesale market energy prices become extremely volatile during hot summer weather when electricity usage spikes. However, consumers are shielded from volatility in wholesale energy prices through a fixed-price retail rate causing demand to be unresponsive to changes in wholesale energy prices. Hence, demand is inelastic in the wholesale energy market resulting in higher overall energy costs. On the other hand, increasing price elasticity through price-responsive demand increases economic surplus relative to uniform rates by avoiding over-consuming when wholesale prices exceed the uniform retail rate or under-consuming when wholesale prices are below the uniform retail rate.
- A uniform rate creates cross subsidies and makes it difficult to recover infrastructure costs without distorting incentives. A uniform rate charges customers that consume most of their energy during low-cost, off-peak periods the same price as those who consume most of their energy during high-cost, peak periods. By charging a fixed-price uniform rate across a broad range of customers each with different preferences and load profiles, some customers end up subsidizing the cost of serving other customers (e.g., users who consume off-peak subsidize

on peak users). Finally, should infrastructure costs (e.g., capacity and transmission) that are driven by peak load also be charged through a fixed-price uniform rate (in ¢/kWh), the incentive to avoid peak usage and encourage off-peak usage is mitigated by spreading such costs evenly across all hours.

- Uniform rates are a disincentive to Market Participants in the supply/delivery chain to encourage demand response. There are also institutional reasons why uniform rates are a barrier to price-responsive demand. First, since price-responsive demand is generally perceived to reduce retail revenue by reducing energy throughput, a uniform rate creates disincentives for entities that are part of the supply/delivery chain to promote demand response.¹¹ Decoupling retail revenue from consumption levels has been introduced as a solution to this disincentive. Second, since the uniform rate is generally coupled with a full-requirements supply contract that offers retail customers a hedge against price volatility, it 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. California Public Utility Commission's landmark decision in 2008 to adopt dynamic retail pricing as default rate for all customer classes is generally viewed as a positive step toward achieving price-responsive demand.

Second, in the electricity sector, price-responsive demand depends on advanced meters and the infrastructure that enables demand response, including metering, communications, control, billing and demand management tools. If a load serving entity wants to bill customers for time-differentiated rates, interval meters are needed at the very least to record customer usage on an hourly basis. Price-response demand requires communications technology that sends hourly price and consumption information in real time to the consumer or their energy-using devices (to determine the level of response), and to billing entities with the capacity to efficiently process hourly price and consumption information from numerous customers. To manage air conditioning usage, control technologies such as smart thermostats that adjust room temperatures

¹¹ Because many demand response programs have been oriented toward reducing peak demand, it is perceived that overall sales would decrease as a result of price response. However, price response that improves system utilization by increasing off-peak usage may maintain overall system throughput and mitigate such perceived revenue erosion.

automatically in response to price signals would greatly facilitate price-response. During shortage events, the system operator needs a secure, dispatch and communication network to control customer loads in response to system emergencies or price signals. 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.

The technical barrier caused by the lack of advanced metering infrastructure is largely due to the economies of scale and scope for the investment in the technological infrastructure needed to facilitate price-responsiveness. Moreover, price-responsive demand tends to reduce energy prices during peak load period yielding external benefits to other customers. As a result, reliance on pure private investment to finance the development of advanced metering infrastructure seems insufficient. Evidently, only large industrial and commercial customers can justify the expense of metering, communications, and enabling technologies at this time. As a result, price-responsive demand has been limited to date.

Given the institutional and technical barriers – default uniform retail rate and lack of advanced metering infrastructure – consumers are likely to stay on fixed-price uniform rate service. This creates two fundamental market deficiencies that prevent the market from realizing the benefits of price-responsive demand described above. First, since consumers cannot respond to real-time prices, real-time demand is inelastic and is susceptible to the potential exercise of market power by producers especially during peak hours when available capacity is scarce. Second, when a utility or load serving entity cannot charge its customers for their incremental costs of services in real-time, and customers are billed for their aggregate energy consumption rather than their hourly energy consumption, inefficiencies result – over-consumption during peak-load periods requiring extra investment in generation and transmission capacity, under-utilization of idle capacity during off-peak periods, and ultimately, higher retail electricity prices and bills.

Uniform Retail Rate in the Shadow of Obsolescence

When all consumers are able to respond to real-time prices, we contend that the average-cost based uniform retail rate is likely to become obsolete. For example, as will be shown below, when consumers are given the opportunity to choose freely between two pricing approaches, uniform pricing and demand-profile pricing, no consumer will stay with uniform pricing in equilibrium. The basic model can be extended to explicitly include demand-profile pricing and real-time pricing approaches. The model extension, which is fairly straightforward, is included in the Appendix for completeness.

Under demand-profile pricing, each consumer pays for the service at a fixed price based on the real-time prices weighted by the demand profile for the individual consumer. For demand-profile pricing to be viable, consumers need to have advanced meters; otherwise retail service provider will suffer adverse selection and moral hazard problems due to the asymmetric information from not being able to track down the actual usage. (Joskow and Tirole, 2006) While demand-profile pricing can be revenue neutral by passing through the real-time prices directly to customers, it can be implemented in such a way that a retailer may include a risk premium and offer a price hedge for some risk-averse consumers. Therefore, when we compare demand profile pricing with uniform pricing, we may abstract from the risk management considerations.

Demand-profile pricing strategically dominates uniform pricing in the sense that in equilibrium, all rational consumers should switch from uniform pricing to demand-profile pricing. This can be established by simulating consumer choice over time. Initially, consumers with the lowest-cost demand-profile are subsidizing the higher-cost consumers under uniform pricing. These consumers should be the first ones who will switch to demand-profile pricing. After the lower-cost consumers have switched to demand-profile pricing, the cost of service for the remaining consumers will increase, and the uniform rate needs to be adjusted upward accordingly. This process feeds itself, including more consumers to switch. This process will continue until the inherent cross-subsidization in the uniform rate is eliminated. Ultimately, no consumer will stay with uniform pricing. This process can be expedited by regulatory reform

making dynamic pricing the default service rather than the uniform flat rate. The above observations are demonstrated in the following result.

Result 3.1 When consumers can choose free between demand profile pricing and uniform pricing, all consumers will ultimately switch to demand profile pricing.

Proof:

Under demand profile pricing, consumer l pays a flat rate, $\hat{P}_l = E[\rho_{l\theta} P_\theta \varphi_{l\theta}] / E[\rho_{l\theta} \varphi_{l\theta}]$. Under uniform pricing, each consumer is charged the price, $\hat{P} = E[\rho_{i\theta} P_\theta D'_{i\theta}(\hat{P})] / E[\rho_{i\theta} D'_{i\theta}(\hat{P})]$. The consumer prefers demand profile pricing, if $\hat{P}_l \leq \hat{P}$, given that the rationing plan is not affected by this choice:

$$\begin{aligned} & E[\text{Max}_{q_{l\theta}} \{\rho_{l\theta} U_{l\theta}(q_{l\theta}) - \hat{P}_l \rho_{l\theta} q_{l\theta}\}] \\ & \leq E[\text{Max}_{q_{l\theta}} \{\rho_{l\theta} U_{l\theta}(q_{l\theta}) - \hat{P}_l \rho_{l\theta} q_{l\theta}\}] \end{aligned}$$

This implies that consumers will choose to stay with uniform pricing only if $\hat{P}_l > \hat{P}$. But we show below that this leads to a contradiction in equilibrium.

$$\begin{aligned} \hat{P} &= \frac{E[\sum_{l \in I} \rho_{l\theta} P_\theta \varphi_{l\theta} D'_l(\hat{P})]}{E[\sum_{l \in I} \rho_{l\theta} \varphi_{l\theta} D'_l(\hat{P})]} = \frac{E[\sum_{l \in I} \rho_{l\theta} \hat{P}_l \varphi_{l\theta} D'_l(\hat{P})]}{E[\sum_{l \in I} \rho_{l\theta} \varphi_{l\theta} D'_l(\hat{P})]} \\ &> \frac{E[\sum_{l \in I} \rho_{l\theta} \hat{P} \varphi_{l\theta} D'_l(\hat{P})]}{E[\sum_{l \in I} \rho_{l\theta} \varphi_{l\theta} D'_l(\hat{P})]} = \hat{P}. \end{aligned}$$

Therefore, no consumer can be found to stay with uniform pricing in equilibrium.

Q.E.D

Suppose that a retail service provider (RSP) purchases energy from the wholesale market based on consumers' demand profiles. Let \hat{p}_l denote the flat retail rate that the RSP charges consumer l . The retail profit is given by $E[\rho_{l\theta} (\hat{p}_l - P_\theta) \varphi_{l\theta} D(\hat{p}_l)]$. The free entry condition implies that

the competitive retail price equals the optimal demand-profile price:

$\hat{P}_l = E[\rho_{l\theta} P_\theta \varphi_{l\theta}] / E[\rho_{l\theta} \varphi_{l\theta}]$. By contrast, Borenstein and Holland (2002) showed that the free entry condition was violated under uniform pricing and concluded that retail competition is inefficient. We summarize the above observations as follows.

Result 3.2: Under demand-profile pricing, retail competition is efficient.

The following result shows that real-time pricing is superior because real-time pricing offers an informational advantage over demand-profile pricing derived from the option value associated with dynamic decision-making.

Result 3.3 For risk-neutral consumers, real-time pricing is superior to demand-profile pricing.

Proof:

First, recall that $D_{l\theta}(\hat{P}_l)$ solves the following maximization problem for consumer l under demand profile pricing:

$$\begin{aligned} & E[\rho_{l\theta} U_{l\theta}(D_{l\theta}(\hat{P}_l)) - \rho_{l\theta} \hat{P}_l D_{l\theta}(\hat{P}_l)] \\ & \equiv E[\text{Max}_{q_{l\theta}} \{\rho_{l\theta} U_{l\theta}(q_{l\theta}) - \rho_{l\theta} \hat{P}_l q_{l\theta}\}]. \end{aligned}$$

Since $D_{l\theta}(\hat{P}_l) = \varphi_{l\theta} D_l(\hat{P}_l)$, we have,

$$\hat{P}_l = \frac{E[\rho_{l\theta} P_\theta \varphi_{l\theta}]}{E[\rho_{l\theta} \varphi_{l\theta}]} = \frac{E[\rho_{l\theta} P_\theta D_{l\theta}(\hat{P}_l)]}{E[\rho_{l\theta} D_{l\theta}(\hat{P}_l)]}.$$

Therefore, for $l \in L$, we obtain,

$$\begin{aligned} & E[U_{l\theta}(D_{l\theta}(P_\theta)) - P_\theta D_{l\theta}(P_\theta)] \\ & \equiv E[\text{Max}_{q_{l\theta}} \{U_{l\theta}(q_{l\theta}) - P_\theta q_{l\theta}\}] \\ & \geq E[\rho_{l\theta} U_{l\theta}(D_{l\theta}(\hat{P}_l)) - \rho_{l\theta} P_\theta D_{l\theta}(\hat{P}_l)] \\ & = E[\rho_{l\theta} U_{l\theta}(D_{l\theta}(\hat{P}_l)) - \rho_{l\theta} \hat{P}_l D_{l\theta}(\hat{P}_l)] \end{aligned}$$

4. Demand Response Incentive and the Use of Customer Baseline

In this section, we address the incentive programs for demand reduction. Policy makers have recognized many of the benefits and barriers to price-responsive demand. Regulators have approved demand-response programs that involve paying customers to reduce or eliminate their demand during high price periods. These programs emphasize demand reduction during peak periods or tight system conditions when the benefit is expected to be the greatest.

Demand reduction is conceptualized as a resource that can be treated like a supply resource. This ‘resource’ perspective is expedient in restructured markets, because it provides retail customers incentives to reduce consumption when the wholesale price is high or the system condition is tight but otherwise retain the access to the retail tariff. However, in the long run after the barriers to price responsive demand discussed above are removed, these incentive programs should no longer be needed. Demand and supply should naturally adjust to market prices efficiently without the need to pay someone extra incentive not to buy the product. As prices for a product that consumers pay begin to increase, consumers will naturally conserve usage of that product. Consumers could use long-term contracts to hedge the price risks. Economic efficiency is advanced when the retail rates reflect wholesale prices.

Since many of these programs pay customers based on how much demand they reduced, it is necessary to estimate what their demand would have been had it not been reduced. This estimated amount is called the Customer Baseline. As we describe below, there are theoretical and practical problems with calculating a Customer Baseline that make it possible for customers to get paid for load that was not actually reduced. These problems call into question the viability of approaches to promote price-responsive demand that rely upon an estimated Customer Baseline. Moreover, if demand is paid to reduce consumption, it may artificially lower prices

and result in prices that are not high enough to retain existing supply, or to attract new supply needed to meet future demand.

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. For the purpose of settlement and billing, a methodology must be implemented that estimates the Customer Baseline from which a payment for reduced consumption can be determined. However, the Customer Baseline is conjectural and is not directly observable and is generally estimated from data that represent the past customer behavior using statistical estimation methods.

A customer does not typically pay for energy associated with the Customer Baseline nor have a financial obligation to purchase the Customer Baseline amount. The ability of customers to consume any quantity of electricity without making a financial commitment before-hand to purchase that quantity enables the customer to sell a quantity of energy that it does not own. The lack of a financial commitment to purchase energy at the Customer Baseline level means that the customer has an ability to increase its revenues from the demand response program by altering its baseline so that it appears to be as high as possible. The incentives to increase the baseline fall into two categories well studied by economists: adverse selection and moral hazard.

The adverse selection problem arises from asymmetrical information on Customer Baseline. Since the baseline is not directly observable, customers usually have better information on their baseline consumption levels than the ISO and can use this information to their advantage in its decision to join a price-response program. Therefore, the program is likely to attract disproportionate participation from customers who anticipate lower consumption for reasons having nothing to do with the incentives paid by the program to reduce load. For instance, if last year’s or last season’s consumption is used as the basis for the Customer

Baseline, firms whose production have shrunk since that time are likely to sign up.¹² The ISO, therefore, could end up paying for load reduction that would have occurred anyway, which has nothing to do with the demand-response program incentives. 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 initially to create a higher baseline. As indicated in an ISO filing with the Commission in February 2008 concerning the Day-Ahead Load Response Program ("DALRP"), some DALRP participants appeared to have artificially inflated their Customer Baselines in order to collect energy and capacity payments without actually reducing load. For example, customers with base-load, 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.¹³

Given that customers alone know their true demand, it is impossible to foresee all possible gaming strategies and the costs they could eventually impose on other customers. More sophisticated measurement and verification methods for estimating the Customer Baseline may help, but will not solve the problem because the problem is the result of the incentive structure created by reliance on an estimated Customer Baseline and not in the Customer Baseline estimation methodology.

Finally, paying customers not to consume energy can cause inefficient price formation in the wholesale energy markets because it can result in customers not consuming power when the value of consumption exceeds the cost of producing the energy. This can happen when the sum

¹² For example, in the fall of 2007, the ISO New England detected an unusual surge in participation in its Day-Ahead Load Response Program. It was found that some of the Customer Baselines were frozen in at higher summer levels, so when demand started naturally decreasing during the off-peak fall season, payments were being made for load reductions that would have occurred in any case. See ISO New England, Docket No. ER08-538-000 (February 5, 2008).

¹³ *Id.*

of the customer's of bill savings and the program's incentive payments, exceed the cost of production. A simple example can be used to illustrate an inefficient market result caused by paying for demand reductions:

Suppose that a consumer is willing to pay the retailer up to \$250/MWh for energy (imagining that this could be the opportunity cost to run an on-site backup generator including time, materials and inconvenience).¹⁴ When the retail rate is \$120 per MWh, the consumer would be willing to offer load reduction at \$130/MWh (= \$250 - \$120). If the real-time LMP of \$140 per MWh, and the load reduction offer clears, it would displace a marginal generating unit offered at \$140 per MWh. This results in a net societal cost increase of \$110 (=250 - 140) per MWh. In other words, the double payment from the demand response program caused a higher cost resource to be dispatched – i.e., the on-site generator at \$250/MWh – while a less expensive resource – the system generator at \$140/MWh – was not utilized. It is possible that the dispatch of the on-site generator caused a reduction in the market clearing price – say from \$150 to \$140/MWh. Even in that situation, society ends up paying a resource cost of \$250/MWh while a substantially less expensive alternative (a system generator at \$150/MWh) went unused.

To further illustrate, suppose that demand reduction receives an incentive payment, $P_t(CBL - Q_t)$, which equals the product of the real-time price (RTP), P_t , and amount of demand reduction from the Customer Baseline level ($CBL = \bar{Q}_0$) to the actual demand, Q_t . To see the effects of the incentive payment on consumer behavior, we write the consumer's surplus as follows:

$$\max_{Q_t} E\{U(Q_t) - P_0 Q_t + P_t(\bar{Q}_0 - Q_t)\}$$

The consumer's surplus is maximized by setting the marginal willing-to-pay to the sum of retail and wholesale prices: $U'(Q_t) = P_0 + P_t$, where P_0 is the fixed retail rate (RR).

Therefore, the incentive payment causes excessive demand reduction that affects both peak and

¹⁴ The use of an on-site generation was used in this example for illustrative purposes only. This example is generally applicable even in situations where 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 a manufacturing business in which it is no longer profitable to operate if the cost of electricity exceeds \$250/MWh – hence, the consumer will cease electricity consumption at prices exceeding \$250/MWh,

off-peak periods. The actual consumption is set at a point on the demand function where the price equals to the sum of retail and wholesale prices: $Q_t = D_t(P_0 + P_t)$. The result suggests inefficient price formation. Figure 1 shows the effect on consumption of a demand response program that includes such excessive compensation – i.e., compensation for demand reduction that includes both customer bill savings and the incentive payments for the same reduction in energy consumption. With the incentive payment, a consumer’s opportunity cost for demand reduction is the difference between the consumer’s willingness-to-pay (denoted by the demand curve) and its retail rate (P_0). When the consumer’s opportunity cost equals the supply curve (which occurs at P_1 on Figure 1), the effect on consumption is as if the price paid by the consumer were the retail rate plus the wholesale market price. As a result, demand is reduced from the baseline level, $CBL_{Peak} (CBL_{Off-peak})$, to $Q_{Peak} (Q_{Off-peak})$, below the efficient level of price-responsive demand for peak as well as off-peak period. The incentive is intended to eliminate excessive consumption during peak periods, but it overshoots, resulting in under-consumption in all periods. As shown later in this report, the social welfare losses – i.e., the reduction in economic surplus – equals the sum of the shaded areas in the Figure 1. Moreover, the total payments for demand response incentive could be substantial, and it could result in a higher retail rate to the disadvantage of the final consumers.

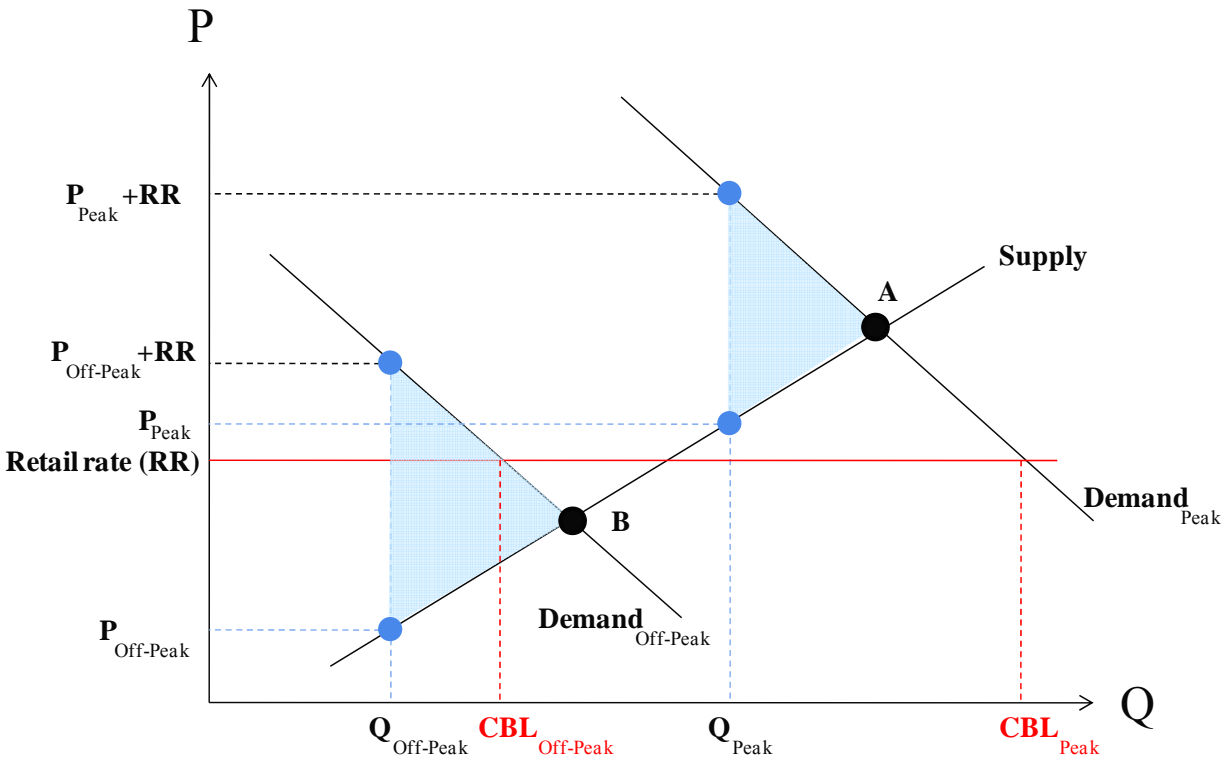


Figure 1. Excessive demand reduction payment causes inefficient price formation

The root cause of the issues created by use of a Customer Baseline in price-response programs – i.e., adverse selection, moral hazard, and price-formation – is that the customer is not obligated to purchase the energy associated with baseline amount. In principle, if a consumer owns the title of a specific amount of energy that is not consumed, there is no issue with the consumer selling demand reduction into the market, which is equivalent to selling the energy to which they have title but do not consume. For example, suppose that a retail customer has a take-or-pay contract with an independent power producer for a fixed amount of energy delivered at a specified time. In this situation, the customer should have the right to sell any unused energy associated with the contract as demand reduction into the real-time market.¹⁵ In other words, if

¹⁵ A load serving entity can act on behalf of the retail customer and the independent generator by entering this transaction into the wholesale market through self-scheduling to simultaneously buy and sell (as price takers) a fixed

the customer purchased and owned the baseline amount, then these issues would not exist and the customer would be able to legitimately “sell-back” the energy to the market at wholesale market clearing prices. The ability of a customer to sell back energy that it does not own is a consequence of the implicit full-requirement contract that is still part of the regulatory regime underlying the fixed-price uniform retail rate tariff.

The “double payments” problem is not new. It has been known to cause deviation from least-cost planning in a vertically integrated utility environment. (Cicchetti and Hogan 1989) In restructured competitive markets, the ambiguity in the regulatory philosophy concerning a consumer’s entitlement to the energy that is not consumed can be conveniently interpreted in different ways leading to inefficient outcomes. Under a full-requirement contract for retail services, consumers are entitled to consume any amount of energy. Therefore, consumers may be led by the implicit assumption that they could resell what they do not consume as demand reductions in the wholesale market. However, the full requirement contract is a volumetric option contract, which is different from a regular fixed-quantity contract for specific amounts of energy under specified terms. While a full-requirements contract may give a consumer a fixed price insurance and the right to buy or consume an unspecified amount up to a capacity level at the fixed price, it does not automatically grant the ownership of any specific amount of energy that has not been bought. Since a consumer is not financially obligated to pay for what has not been consumed even after the demand reduction is cleared in the wholesale market, double payments provide an excessive incentive for demand reduction.

One possible approach to resolve this problem is to define the Customer Baseline as a quantity of energy for which the customer has procured and owns. If a consumer owns the specific amount of energy that is not consumed, there should be nothing preventing the consumer from selling it as demand reduction back to the wholesale market. But it is inefficient to pay consumers for not consuming the energy that they do not own. Property rights and ownership

amount of energy in the day-ahead energy market. Since the bids and offers form a perfect hedge, this transaction is not exposed to the price risks in the wholesale market, if everything goes according to the agreement. The customer should receive a demand-reduction payment only if the actual consumption falls below the contracted amount, and the deviation is compensated at the prevailing real-time price. This is in effect how demand response programs have been implemented in vertically-integrated utilities. The two-part real-time pricing program with Georgia Power is such an example. Although the customer baseline is a transfer pricing agreement rather than a contract, the principle remains the same.

entitlements are essential to a market. Therefore, using an amount of energy for which a consumer has clear title as the Customer Baseline is consistent with the principles of competitive markets. For example, in an organized ISO/RTO market if the Customer Baseline level is established by a contracted quantity through bilateral contracts or purchase obligations in the Day-Ahead Energy Market, demand reductions can be offered into the wholesale market like an offer to supply generation and produce an efficient market outcome devoid of the price formation problem and adverse incentives.¹⁶ In this case, demand reduction is a resale of energy that is not consumed.

Let us denote by (\bar{P}_0, \bar{Q}_0) the contracted price and quantities (as a load profile). In this case, the objective function for the customers can be revised as follows,

$$\max_{Q_t} E\{U(Q_t) - \bar{P}_0 \bar{Q}_0 + P_t(\bar{Q}_0 - Q_t)\}$$

The first-order optimality condition for consumer's decision yields, $U'(Q_t) = P_t$. Indeed, when the contracted amount, \bar{Q}_0 , that the consumer has purchased is used as customer baseline, the efficiency of price formation is restored.

If the contract price is lower than the demand-weighted average of expected wholesale prices, the customer will find a higher customer baseline level attractive and increase the amount in the contract. This tends to raise the contract price moving it closer to the expected wholesale prices. On the other hand, if the contract price is too high, the customer would reduce the demand for contracted energy and lower its price. Therefore, using the contracted amount as the customer baseline is consistent with the outcome of competitive markets. When the customer baseline is determined through competition in the forward market, contractual obligation removes the adverse incentives for the moral hazard and adverse selection problems.

Result 4.1: When the customer baseline level is based on forward contract, demand response incentive leads to efficient price formation.

An Illustrative Example

¹⁶ This is similar to, for example, in Georgia Power's two-part real-time pricing (RTP) Program, in which retail customers can participate in the RTP program and subscribe to a baseline retail Time-of-use (TOU) rate.

The following example shows how extra incentives can reduce consumer benefits in the long term. Consider an electricity system with a peak load period for twenty-five percent of the hours and an off-peak period for seventy-five percent of the hours, assuming that the demand is constant within each period. The hourly demand functions for electric energy in these periods varies linearly with the price paid by the consumer as follows:

$$\text{Peak-load period demand (MW): } D_{Peak}(P) = 20,000 - 20 \times P$$

$$\text{Off-peak period demand (MW): } D_{Off-Peak}(P) = 10,000 - 10 \times P$$

Assume that all customers pay a uniform fixed retail rate, which is set such that total retail revenues equal generator revenues. Further assume that there is a single type of generator with an operating cost of \$60 per MWh and an amortized capital cost of \$100 per MWh, which can be recovered through scarcity pricing in the wholesale energy market during the peak period. Thus, the wholesale market price would be \$60 off-peak and \$160/MWh on-peak, the uniform fixed retail rate is \$100/MWh, the peak load is 18,000 MW and the off-peak load is 9,000 MW.

Now consider Case 2, in which 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 7% to 16,800 MW and rise above 9,000 MW during the off-peak period by 4% to 9,400 MW. This improves the economic efficiency and increases the consumers' surplus by \$60,000.

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 3,200 MW of demand reduction, which represents an 18% reduction from the Customer Baseline. 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. Such a price drop would cause large transfer benefits during the peak period as bill savings for all customers. But such a situation is unsustainable because the generation cannot remain as

a commercially viable business 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) 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 \$113/MWh. With the extra incentive, the demand responds to the sum of wholesale and retail prices, which equals \$273/MWh during the peak period. As a result, the peak demand for this case is reduced to 14,546, or 19% lower than the level in the uniform pricing case and 13% lower than the optimal level in the real-time pricing case. Even more strikingly, the off-peak demand is 8% lower than the uniform pricing case and 12% 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 1, the demand-as-supply resource would reduce both the total economic surplus *and the consumer surplus* by an amount which is more than four times larger than the economic benefit from real-time pricing. By contrast, if demand is responsive to spot prices that exceed the retail rate, but *without* the extra incentive of avoiding retail payments while receiving the LMP (Case 4), economic surplus and consumer surplus increase by 60% as much as when demand is always responsive (Case 2).

In summary, this example illustrates that the excessive incentive results in inefficient demand reduction and negative consumer benefits in the long term.¹⁷

Table 1. Summary of Illustrative Results

¹⁷ The important insight is consistent with those in Chao (2008) that price response demand with the extra incentive of double payments (Case 3) is substantially less efficient than uniform pricing. The welfare loss from inducing under-consumption in every period more than offsets the welfare gain from avoiding excessive consumption during peak periods. 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 as suppliers adjust their investment/retirement decisions, as discussed above.

	Case 1: Demand Insensitive to Spot prices	Case 2: Demand Responsive to Spot prices	Case 3: Demand Responsive with Extra Incentive	Case 4: Demand Responsive when Spot price >Retail Rate
Retail Rate	\$100/MWh	\$97/MWh	\$113/MWh	\$100/MWh
Peak Load	18,000 MW	16,800 MW	14,546 MW	16,800 MW
Off-Peak Load	9,000 MW	9,400 MW	8,273 MW	9,000 MW
Economic Surplus and Consumer Surplus ¹⁸	0	\$360,000/Day	-\$1,545,000/Day	\$216,000/Day

5. CONCLUSION

Price-responsive demand improves the economic efficiency by discouraging low value energy consumption when real-time wholesale energy prices are high during the peak period and encouraging high value energy consumption when real-time wholesale energy prices are low during the off-peak period. The default uniform retail rates and the inadequate advanced metering and demand management infrastructure prevent full potentials of price-responsive demand from being realized.

Before these technical and institutional barriers are removed, demand response incentive could improve economic efficiency within a hybrid market structure. However, the use of customer baseline as the basis for load reduction payments is susceptible to gaming problems, and excessive demand-response incentives can cause inefficient price formation and excessive

¹⁸ In this example, the total economic surplus equals consumer’s surplus because the producer’s surplus is zero due to the assumption that there is only one generation technology. See Appendix B for details.

demand reduction to the disadvantage of final consumers. A contract-based customer baseline appears to offer an efficient remedy. The economic framework discussed in this paper could offer a basis to facilitate economic evaluation of alternative approaches.

Appendix A: An Integrated Model

In this section, we extend the basic model to incorporate three alternative pricing mechanisms: real-time pricing, demand-profile pricing and uniform pricing. We associate price-responsive demand with real-time pricing, price-sensitive demand with demand-profile pricing and price-insensitive demand with uniform pricing. According we let R , S and I be subsets of L that represent, respectively, price-responsive demand (real-time pricing), price-sensitive demand (demand-profile pricing) and price-insensitive demand (uniform pricing).

In the extended model, the social planner's problem is to choose a set of retail prices, including real-time prices, $(p_{r\theta})$ demand-profile prices (\hat{p}_l) and a uniform price (\hat{p}) , supply prices (P_θ) , rationing plan $(\rho_{r\theta}, \rho_{l\theta}, \rho_{i\theta})$ and investment plan $(s(c))$ that maximize the social welfare function:

$$\begin{aligned} \text{Max} \left\{ E \left[\sum_{r \in R} \rho_{r\theta} U_{r\theta}(D_{r\theta}(p_{r\theta})) + \sum_{s \in S} \rho_{s\theta} U_{s\theta}(D_{s\theta}(\hat{p}_s)) + \sum_{i \in I} \rho_{i\theta} U_{i\theta}(D_{i\theta}(\hat{p})) \right. \right. \\ \left. \left. - C_\theta(S_\theta(P_\theta)) - \int_{\underline{c}}^{\bar{c}} k(c) \alpha_\theta(c) s(c) dc \right] \right\} \end{aligned}$$

subject to

$$\begin{aligned} S_\theta(P_\theta) &\geq \sum_{r \in R} \rho_{r\theta} D_{r\theta}(p_{r\theta}) + \sum_{s \in S} \rho_{s\theta} D_{s\theta}(\hat{p}_s) + \sum_{i \in I} \rho_{i\theta} D_{i\theta}(\hat{p}) \\ \mathbf{1} &\geq (\rho_{r\theta}, \rho_{s\theta}, \rho_{i\theta}) \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{1}$ and $\mathbf{0}$ denote vectors with components of 1's and 0's, respectively.

The first order optimality conditions can be written as follows.

Optimal pricing

For $r \in R$ and $s \in S$,

$$P_{r\theta} = P_\theta = \begin{cases} C'_\theta(S_\theta(P_\theta)), & \text{if } \rho_\theta = 1 \\ \text{VOLL}_\theta, & \text{if } \rho_\theta < 1 \end{cases}$$

$$\hat{P}_s = \frac{E[\rho_{s\theta} P_\theta D'_{s\theta}(\hat{P}_s)]}{E[\rho_{s\theta} D'_{s\theta}(\hat{P}_s)]} = \frac{E[\rho_{s\theta} P_\theta \varphi_{s\theta}]}{E[\rho_{s\theta} \varphi_{s\theta}]},$$

$$\hat{P} = \frac{E[P_\theta \sum_{i \in I} \rho_{i\theta} D'_{i\theta}(\hat{P})]}{E[\sum_{i \in I} \rho_{i\theta} D'_{i\theta}(\hat{P})]}.$$

Optimal investment

$$k(c) = E[\alpha_\theta(c)(P_\theta - c)^+], \quad \text{if } s(c) > 0.$$

$$LOLP = Pr \left\{ S_\theta(\bar{c}) < \sum_{l \in L} D_{l\theta}(P_{l\theta}) \right\} = \frac{k(\bar{c})}{VOLL - \bar{c}}$$

where

$$P_{l\theta} = \begin{cases} P_\theta, & \text{if } l \in R \\ \hat{P}_l, & \text{if } l \in S. \\ \hat{P}, & \text{if } l \in I \end{cases}$$

Optimal rationing

$$\rho_{r\theta} = 1$$

With a mixture of price-responsive, price-sensitive (but not price responsive) and price-insensitive consumers, the price-responsive demand would respond to rising real-time price with demand reduction until the price reaches the shortage price, $VOLL_\theta$, while the price-sensitive (demand-profile pricing) and price-insensitive (uniform pricing) consumers are not price-elastic in the real-time wholesale market. When the real-time price reaches the scarcity price, $VOLL_\theta$, the non-price-responsive demand will be curtailed during the shortage event while the price-responsive demand will continue to respond to real-time prices and no rationing is imposed on price response demand. Only after the entire price-insensitive demand is reduced to zero, the real-time price may rise again above the scarcity price, $VOLL_\theta$, and the price-responsive demand would drop accordingly. Results 3.1 – 3.4 are generalized as follows:

Result A.1: The price-responsive demand is charged the competitive wholesale market price, which equals the marginal cost of production, during a normal period ($\rho_\theta = 1$), and the marginal value of lost load, during a shortage period ($\rho_\theta < 1$), i.e.,

$$P_\theta = \begin{cases} C'_\theta(S_\theta(P_\theta)), & \text{if } \rho_\theta = 1 \\ VOLL_\theta, & \text{if } \rho_\theta < 1 \end{cases}$$

Result A.2.1 The optimal demand-profile retail price equals to the expected demand-profile-weighted average of competitive wholesale market prices,

$$\hat{P}_s = \frac{E[\rho_{s\theta} P_\theta \varphi_{s\theta}]}{E[\rho_{s\theta} \varphi_{s\theta}]}$$

Result A.2.2 The optimal uniform retail price equals to the expected marginal-demand-weighted average of competitive wholesale market prices,

$$\hat{P} = \frac{E[P_\theta \sum_{i \in I} \rho_{i\theta} D'_{i\theta}(\hat{P})]}{E[\sum_{i \in I} \rho_{i\theta} D'_{i\theta}(\hat{P})]}.$$

Result A.3 The efficient investment plan is characterized by the zero-profit free-entry condition.

$$k(c) = E[\alpha_\theta(c)(P_\theta - c)^+], \quad \text{if } s(c) > 0.$$

Result A.4 The optimal capacity level is attained when the lost of load probability equals

$$LOLP = k(\bar{c}) / (VOLL - \bar{c}).$$

Appendix B

Assumptions and Initial Set-up:

$$\text{Peak period demand (MW):} \quad Q_{Peak}(P) = 20,000 - 20 \times P$$

$$\text{Off-peak period demand (MW):} \quad Q_{Off-Peak}(P) = 10,000 - 10 \times P$$

Therefore, inverse demand functions are:

$$P(Q_{Peak}) = 1,000 - 0.05 \times Q_{Peak}$$

$$P(Q_{Off-Peak}) = 1,000 - 0.1 \times Q_{Off-Peak}$$

In addition:

Peak hours are 25% of all hours or 6 hours during a day

Off-peak hours are 75% of all hours or 18 hours during a day

On-peak LMP market price (P_{Peak}) is \$160/MWh

Off-peak LMP market price ($P_{Off-Peak}$) is \$60/MWh

Retail price is always set to ensure that total retail revenues equal generator revenues.

CASE 1: All demand is insensitive to wholesale price

In this case, all customers are on fixed-price uniform rate and always pay the retail price (in both peak and off-peak hours) – hence, to find the retail price, we use the inverse demand equations substituting RR (the retail rate) for $P(Q_{Peak})$ and $P(Q_{Off-Peak})$:

$$1,000 - 0.05 \times Q_{Peak} = 1,000 - 0.1 \times Q_{Off-Peak} \Rightarrow Q_{Peak} = 2Q_{Off-Peak}$$

Next, we substitute the above result in the revenue equivalence expression (recognizing that 0.25 of the time the quantity comes from on-peak hours and 0.75 of the time it comes from off-peak hours) so that:

$$\text{Generator revenue} = \text{Retail revenue}$$

$$(0.25)(Q_{\text{Peak}})(P_{\text{Peak}}) + (0.75)(Q_{\text{Off-peak}})(P_{\text{Off-peak}}) = [(0.25)(Q_{\text{Peak}}) + (0.75)(Q_{\text{Off-peak}})]RR$$

Substituting $P_{\text{Peak}} = 160$, $P_{\text{Off-peak}} = 60$, $Q_{\text{Peak}} = 2Q_{\text{Off-peak}}$, we obtain $RR = \$100$ and

$$Q_{\text{Peak}}(RR) = 20,000 - 20 \times 100 = 18,000 \text{ MW}$$

$$Q_{\text{Off-Peak}}(RR) = 10,000 - 10 \times 100 = 9,000 \text{ MW.}$$

With the assumption of single supply technology, the economic surplus equals the consumer surplus, and the daily economic surplus is calculated as:

$$\begin{aligned} ES &= \frac{1}{2} \sum_t Q_t(RR)[P_t(0) - RR] \\ &= \frac{1}{2}(18,000)(1000 - 100)(0.25 \times 24) + \frac{1}{2}(9,000)(1000 - 100)(0.75 \times 24) \\ &= \$121,500,000 \end{aligned}$$

CASE 2: All demand is responsive to wholesale real-time prices

In this case, To find the quantities consumed, we substitute the on-peak on off-peak prices into the corresponding demand equations:

$$Q_{\text{Peak}}(P_{\text{Peak}}) = 20,000 - 20 \times 160 = 16,800 \text{ MW}$$

$$Q_{\text{Off-Peak}}(P_{\text{Off-peak}}) = 10,000 - 10 \times 60 = 9,400 \text{ MW}$$

The retail rate has to satisfy the revenue equivalence requirement—hence, it equals:

$$RR = \frac{\sum_t P_t Q_t}{\sum_t Q_t} = \$97.33 / MWh,$$

where Q_t and P_t are the relevant hourly quantities (peak or off-peak) and the relevant hourly prices (peak or off-peak).

Since in Case 2 all demand is responsive to the LMP price, economic surplus is given by:

$$\begin{aligned} ES &= \frac{1}{2} \sum_t Q_t (P_t) [P_t(0) - P_t] \\ &= \frac{1}{2} (16,800)(1000 - 160)(0.25 \times 24) + \frac{1}{2} (9,400)(1000 - 60)(0.75 \times 24) \\ &= \$121,860,000 \end{aligned}$$

In comparison with case 1, the economic surplus for case 2 is increased by \$360,000 per day.

CASE 3: All demand is price responsive after provided with an extra incentive payment to reduce demand such that the customer saves its retail rate and is paid the full wholesale real-time price for demand reductions

In this case, the consumers' first-order condition requires that they consume as if the price were the sum of retail and wholesale prices, $RR+P_t$. In other words, we have to find $Q_{Off-peak}(RR+P_t)$ and $Q_{peak}(RR+P_t)$ for every hour (peak and off peak). In addition, the process requires a number of iterations until the retail price value converges to a stable value (up to a user-specified threshold). The steps are the following:

1. Start with an initial retail price $RR = \$100$
2. Find the hourly actual quantities and the “new” baselines: $Q_{peak}(RR)$ and $Q_{Off-peak}(RR)$
3. Recalculate the retail price RR to include the incentive payment based on the new baselines
4. Repeat Steps 1-3 until the change in the retail price RR is smaller than the threshold level

Step 3 requires re-calculation of retail price, RR , because the incentive payments create a feedback into the retail price. While the final consumers only pay for the actual quantity consumed at the “new” retail price, some consumers receive an incentive payment equal to the difference between their (“new”) baseline and the actual quantity consumed times the wholesale price in that hour. The revenue equivalence still has to hold: retail revenue = generator revenue.

$$\text{Retail revenue} = \sum_t RR \times Q_t(RR + P_t) - \sum_t P_t [Q_t(RR) - Q_t(RR + P_t)]$$

$$\text{Generator revenue} = \sum_t P_t Q_t(RR + P_t)$$

$$\text{Therefore, } RR = \frac{\sum_t P_t Q_t(RR)}{\sum_t Q_t(RR + P_t)}$$

The above expression is used in step 3 in order to calculate the “new” retail rate.

Once we have obtained the converged value for RR , we need to calculate the economic surplus for each hour in the following way:

$$\begin{aligned} ES &= \frac{1}{2} \sum_t Q_t(RR + P_t) [P_t(0) - P_t - RR] + \sum_t Q_t(RR) P_t \\ &= \frac{1}{2} (14,546)(1000 - 160 - 112.70)(0.25 \times 24) + \frac{1}{2} (8,273)(1000 - 60 - 112.70)(0.75 \times 24) \\ &\quad + (17,746)(160)(0.25 \times 24) + (8,873)(60)(0.75 \times 24) \\ &= \$119,955,000 \end{aligned}$$

CASE 4: All demand is price responsive when the wholesale price exceeds the retail rate

In this case, the customer receives an incentive equal to the difference between the wholesale price and the retail price times the customer baseline. However, this incentive materializes only when the wholesale price is greater than the retail price. This occurs only during the peak hours. Therefore, the incentive payment reduces “over-consumption” but not “under-consumption”. The economic surplus is calculated as follows:

$$\begin{aligned}
ES &= \frac{1}{2} \sum_t Q_t(\text{Max}(RR, P_t)) [P_t(0) - \text{Max}(RR, P_t)] + \sum_t Q_t(RR)(P_t - RR)^+ \\
&= \frac{1}{2}(16,800)(1000 - 160)(0.25 \times 24) + \frac{1}{2}(9,000)(1000 - 100)(0.75 \times 24) + (18,000)(160 - 100)(0.25 \times 24) \\
&= \$121,716,000
\end{aligned}$$

References

- Allaz, B. and Vila, J.L. (1993) "Cournot Competition, Forward Markets and Efficiency." *Journal of Economic Theory*, Vol. 59, pp. 1–16.
- Barmack, M., Kahn, E. and Tierney, S (200x), "A Cost-Benefit Assessment of Wholesale Electricity Restructuring and Competition in New England" *Journal of Regulatory Economics*
- Boiteux, M. (1960), "Peak load Pricing," *Journal of Business*, 33:157-79 [Translated from the original in French published in 1951.]
- 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.
- Brown, S. and Sibley, D. (1986) *The Theory of Public Utility Pricing*, Cambridge University Press.
- Carlton, D. (1977) "Peak Load Pricing with Stochastic Demand." *American Economic Review*, 67, 5, 1006-1010.
- Chao, H. P. (1983) "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty" *Bell Journal of Economics*, 14(1), 179-190.
- Chao, H. P. (2008) "An Economic Framework and Evaluation of Price Responsive Demand", presentation to the NEPOOL Market Committee, October 22, <http://www.iso->

ne.com/committees/comm_wkgrps/mrks_comm/mrks/mtrls/2008/oct21222008/index.html
a8_iso_presentation_10_22_2008

- Chao, H. P. 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. P. Oren, S. and Wilson, R. (2005) “Restructured Electricity Markets: A Risk Management Approach” Electric Power Research Institute, Palo Alto, CA.
- Chao, H. P., Oren, S., Smith, S., and Wilson, R. (1986) “Multilevel Demand Subscription Pricing for Electric Power” *Energy Economics* 8: 199-217.
- Chao, H. P. and Wilson R. (1987) “Priority Service: Pricing, Investment, and Market Organization.” *American Economic Review*, Vol. 77, pp. 899–916.
- Chao, H. P. and Wilson R. (2005) “Resource Adequacy and Market Power Mitigation via Option Contracts” In Report no. 1010712, Electric Power Research Institute, October
- Cicchetti, Charles and William Hogan (1989), “Including Unbundled Demand-side Options in Electric Utility Bidding Programs”, *Public Utility Fortnightly*, June 8.
- Crew, M. and Kleindorfer, P. (1976) “Peak Load Pricing with a Diverse Technology”, *Bell Journal of Economics*, Vol. 7(1), pp. 207-231.
- Crew, M. and Kleindorfer, P. (1978) “Reliability and Public Utility Pricing”, *American Economic Review*, Vol. 68, pp. 31-40.
- 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

- Faruqui, A. (2007) “Pricing Programs: Time-of-Use and Real Time”, *Encyclopedia of Energy Engineering and Technology*, Publisher: Taylor & Francis, London, UK.
- 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. (2008) “Household Response to Dynamic Pricing of Electricity: A Survey of Seventeen Experiments”, The Brattle Group Report
- FERC (2006) Demand Response and Advanced Metering, Federal Energy Regulatory Commission Staff Report, Docket AD06-2-000, Washington D.C.
- Holland, S. and Mansur, E. (2006) “The Short-Effects of Time-Varying prices in Competitive Electricity Markets”, *Center for the Study of Energy Markets*, CSEM WP 143R, University of California Energy Institute, Berkeley, CA
- Joskow, P. and Schmalensee (1983) *Markets for Power, An Analysis of Electrical Utility Deregulation*, the MIT Press, Cambridge, MA
- Joskow, P. (2001) “California’s Electricity Crisis”, *Oxford Review of Economic Policy*, Vol. 17, No. 3, 365-388
- Joskow, P. (2006) “Competitive Electricity Markets and Investment in New Generating Capacity”, Working Paper, MIT
- 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
- Ruff, L. (2002) “Economic Principles of Demand Response in Electricity”, Edison Electric Institute, Washington D. C.

Stoft, Steven (2002), *Power System Economics*, IEEE Press.

Telson, M. (1975) “The Economics of Alternative Levels of Reliability for Electric Power Generation Systems”, *Bell Journal of Economics*, Vol. 6, 679-694

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.

Willems, B., Rumiantseva, I., and Weigt, H. (2009) “Cournot versus Supply Functions: What does the data tell us?” *Energy Economics* 31: 38–47

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

Wilson, R. (1993) *Nonlinear Pricing*, Oxford University Press, New York.

