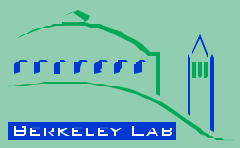
LBNL-52530

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY



Social Welfare Implications of Demand Response Programs in Competitive Electricity Markets

Prepared by
Richard N. Boisvert* and Bernard F. Neenan
Neenan Associates

Prepared for
Charles Goldman

Energy Analysis Department
Ernest Orlando Lawrence Berkeley National Laboratory
University of California Berkeley
Berkeley, California 94720

Environmental Energy Technologies Division

August 2003

http://eetd.lbl.gov/ea/EMS/EMS_pubs.html

* Richard N. Boisvert (rnb2@cornell.edu) is a professor in the Department of Applied Economics and Management, Cornell University and a Senior Academic Associate with Neenan Associates. Bernard Neenan (bneenan@bneenan.com) is President of Neenan Associates.

Work reported here was coordinated by the Consortium for Electric Reliability Technology Solutions (CERTS) and funded by the Assistant Secretary of Energy Efficiency and Renewable Energy, Distributed Energy and Electricity Reliability Program, Transmission Reliability, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Disclaimer

Table of Contents

Figures and Tables

Executive Summary

The price volatility exhibited by wholesale electricity markets has stymied the movement to restructure the industry, and may derail it altogether. Market designers argue that prices are superior to regulation for directing long-term investments to the proper location and function, and that price volatility is a natural manifestation of a robustly competitive market. However, episodes of prices that soar to previously unimaginable heights try customers' patience and cause policy makers to reconsider if the prize is worth the consequences.

As a result, there is growing enthusiasm for short-term demand response (DR) programs, including real time pricing (RTP) service, in regulated and competitive markets and load curtailment programs offered by utilities and ISOs. However, there is not universal agreement as to how such programs should be structured or what entity or entitles should offer them to retail customers. There has been only limited and somewhat abstract discussion in the literature to demonstrate exactly how DR programs can contribute to market efficiency, the management of market risk, and the overall social welfare in restructured electricity markets. This explains in part the current controversy over the role and value of DR in wholesale electricity markets, and who should be implementing them to end use customers

The debate can be summarized as follows: should load curtailments by retail customers be treated as resources that supplement generation reserves and compete against generation energy supply bids, as is currently accommodated to various degrees by the most of the existing ISOs? Or, does this practice amount to the ISO overstepping its charter, resulting in subsidies that disrupt long-term market efficiency more than they contribute in terms of short-term benefits (Ruff, 2002)? Setting the self-interests of the parties to the debate aside, the salient issue is whether the benefits, however distributed, justify incorporating DR directly into the design and structure of wholesale electricity markets. Or, should we stand resolutely on principle, waiting for DR to come about naturally as a result of initiatives by retailers or customers based on their expectations of private benefits (i.e., the value of DR only to their specific portfolio). We contend that at the heart of that issue is whether electricity is a private or public good, as that

determination clarifies whether there is justification for accommodations in wholesale and retail market transactions to attract participation in DR programs, or alternatively to let customer load participation come about in the course of competition, at whatever rate private valuations dictate.

This paper demonstrates that there is a gap between the private and public value of DR that justifies treating load as a resource and paying market prices or value for curtailments undertaken at the direction of the ISO. We focus primarily on two classes of ISO-administered DR programs: energy programs in which retail customers bid load curtailments into ISO-run energy markets and are paid at the market-clearing price if the curtailment is scheduled, and capacity programs in which customer load curtailments are dispatched by an ISO in order to preserve system reliability and are compensated accordingly. These activities are distinguished from demand response implemented by retailers specifically to lower their supply costs or undertaken by customers to reduce their utility bills, in either case with no explicit regard to the system or social consequences.

Our analysis initially focuses on the market for electric energy. We demonstrate that unless some retail customers are directly exposed to dynamic prices (as opposed to paying conventional fixed-rate tariffs), and adjust their demands accordingly, unnecessary and potentially large deadweight social welfare losses will persist indefinitely and limit realization of the value of market restructuring. We show that these welfare losses could be significantly reduced, or eliminated altogether by introducing ISO-implemented DR programs in which at least some customers are paid to reduce load when prices are high. We allow that if somehow the same level of DR were forthcoming by customers based solely on their bill savings, without any additional inducements, the net result would be better by the amount of incentive paid under our scheme. But, if the natural level of DR does not approach the social optimum, then we demonstrate how to quantify what society should be willing to pay to achieve that end.

We demonstrate graphically that the potential welfare gains from energy market DR programs are highest in spot energy markets where both the supply and demand curves are initially extremely price inelastic (e.g., the "steeper" both curves are), which is the case during at least a few hours of the year in every market: some markets have seen

prolonged periods where such conditions prevail.[1] As noted above, this result extends with no loss of generality to customers in RTP or other programs implemented by retailers or utilities that involve no explicit or implicit payment to do so, and that utilize market-clearing prices as incentives. However, realizing the same level of benefits will be difficult in such programs as individual retailers or utilities lack the prescience of the ISO that results from its full-market perspective.[2] Moreover, to attain this result naturally, customers must be inclined to undertake market price risks and have the means to respond cost-effectively. The historic low participation by customers in RTP-type program provides some evidence that the natural response will fall short of the social optimum.

In the case of electricity markets, we argue that the realization of potential welfare gains from DR is further complicated because many customers are unwilling to continually adjust electricity usage, because the transactions costs to them of doing so outweigh the potential benefits. In other words, the bill savings available do not justify the aggravation and costs associated with following hourly prices every day in order to react under seldom-achieved conditions. A small portion of customers may be able to profitably adjust demand to real time prices. For these customers, the 'transactions' costs of demand response may be low because of the nature of how they use electricity, the presence of control technology, or access to on-site generation. These customers would be receptive to services whereby their energy usage was priced at prevailing wholesale market prices, such as those produced by the day-ahead electricity market administered by the ISOs. If the experience with utility RTP programs of the past dozen years is any indication, however, the number of customers that will volunteer for such service is small. Even in their limited application, most of the RTP programs are not what economists initially had in mind--customers facing the market price for any and all usage -- but instead they involve a hedge, using a two-part, revenue-neutral model, that seems

[1] As the supply curve becomes steeper, ceteris paribus, the net welfare benefits move in favor of DR load reduction. Similarly, the less price responsive (steeper) the initial demand curve, the larger are the net welfare gains from a DR load reduction program. However, it does not necessarily follow that private valuations follow suit, and that is the case in the case that electricity is a public good.

[2] Typically, utility RTP program prices are developed by the sponsoring utility to reflect its marginal market position, utilizing synthetically derived marginal outage costs to replicate competitive wholesale market prices. A notable exception is the Niagara Mohawk default rate for large customers that use the NYISO day-ahead prices as the basis for setting hourly energy commodity charges.

to affect both customers' willingness to participate and to adjust electricity use in response to price changes. [3]

Fortunately, there are indications that a third group exists: customers that are willing and able to respond to price, but only on a limited basis.[4] It is through the actions of these customers as participants in DR programs that dead weight losses can be reduced substantially if their DR capability is treated as a resource dispatchable by the ISO with regard to minimizing system costs and customers are paid the market value of their curtailments. It is towards this group of customers that the ISO's DR programs should be designed and directed.

We also extend and expand our analysis to account explicitly for the supply and demand for electricity both as a commodity and as capacity reserves. Reserves are additional, stand-by resources that are acquired and committed for the purposes of maintaining system reliability at acceptable levels. To characterize the value of DR as reserves, our modeling approach incorporates several distinctive features. First, the customer's demand for reserves is modeled as a damage control agent rather than a conventional input to its production process or business activity. This damage control agent reduces both the probability of power outages to the firm, and the firms' economic losses from such outages, by reducing the difference between planned output and output that would actually be realized in the event of an outage. Second, we also recognize explicitly that reserves are a public good in the sense that all customers are provided the same level of reliability due to the common transmission and distribution system.

In economic terms, a public good is non-excludable in the sense that once it (e.g., system reserves) is provided, even those who would fail to pay for it could not be excluded from enjoying the benefits. In this case the public value of system reserves can often be much higher than the private value to any individual customer. Thus, customers,

[3] The revenue-neutral RTP model results in a discontinuity in customer's demand curve. At loads above the level of the CBL, customers are exposed to high prices. But, at the CBL they pay the average tariff rate, regardless of the prevailing price. So, it is not necessarily irrational for a customer to reduce load to its CBL, but no farther, despite the bill saving incentives, an action some RTP participants refer to as hiding behind the demand curve. These customers simply have devised a disjoint demand relationship that may be the result of transactions costs that are higher for giving up base usage (i.e., load below the CBL) than incremental discretionary usage (i.e., load above the CBL).

[4] Utilities have utilized this principle to operate interruptible, curtailable and load control programs for the past 20 years.

acting only in their own self-interests, would not purchase sufficient reserves to maintain system reliability. Although reserves are not often described as public goods, it is precisely for this reason that ISO's are charged with the responsibility of securing sufficient reserves to maintain system reliability.

By treating reserves as a damage control agent and recognizing the public-good nature of system-wide reserves, we are able to differentiate explicitly between the public and private value of a customer's demand for reserves. This valuation gap, which at times is a chasm, is made transparent by modeling five increasingly comprehensive situations within this framework to systematically isolate and characterize the effects of different actions by customers and generators.[5]

We find that once load reductions are explicitly counted as reserves, firms would reduce usage, albeit by a small amount, because they find it to their private advantage to do so increase system reliability, even if they can only capture the "private" benefits from this enhanced system reliability. As one would expect, these independent actions by firms fall well short of the optimal level of load reduction from society's point of view; each firm acts only to their own benefit. However, by explicitly recognizing the public-good nature of reserves explicitly, the value of any single firm's load reduction to all other customers increases, at times dramatically. Such load reduction is worth the combined value of the aggregate reduction in expected outage costs for all firms due to the added reserves. Therefore, unless customers are offered an inducement above their individual private valuation, they will be unwilling to provide sufficient load curtailment, and the full benefit will not be realized. This means that there are opportunities for gains from trade if loads that can reduce under conditions consistent with market operations are compensated by those loads that can not easily curtail, but would gain from their doing so. But, only the central market operator, the ISO or its equivalent, is in the position to manage such trade to its social optimum. The inducement it offers, being paid market

[5] In an initial case, the firm maximizes expected profit for a given level of system reserves. Next, the firm is allowed to reduce load below its customer base load (CBL); this load reduction is recognized by the ISO as an addition to system reserves. In the third case, the public-good nature of the system-wide reserves is considered explicitly. In the fourth and fifth case, the combined load curtailments of firms are assumed to affect electricity price, and generators can supply reserves over and above the fixed level in the initial case, respectively.

prices for the value of the resource provided, enables trades that cannot come about naturally, but that achieve the desired result from the perspective of all parties.

Clearly, only a central market agent responsible for market operations (e.g., an ISO or RTO, or in their absence, the monopoly utility) is capable of the prescience required to achieve the optimal level of DR. By explicitly recognizing the system-wide, social value of additional reserves, there is no inherent subsidy involved in paying for these curtailments, as some have suggested. These are transactions among willing buyers and sellers of reliability or price hedges. When there are sufficient reserves available, the difference in the value of reserves from load reduction and that available from generators is equated to the difference in the additional cost of obtaining reserves from these alternative sources. Under normal system conditions (i.e., sufficient reserves) this difference would be very small. Consequently, reserves provided through load reductions should command no premium in the market relative to reserves supplied from generators, even when the public-good nature of reserves is recognized explicitly.

However, under an emergency situation, where additional generator-supplied reserves are not available to dispatch, the situation is quite different. Under this disequilibrium situation, the loss of load probability (LOLP) for all firms would rise dramatically, as would the proportion of firms' output lost due to the outage, giving rise to a substantial increase in the combined expected outage costs of all firms. The social value of the load reduction, which in this case constitutes the only available source of additional reserves, the benefit of which inures to all electricity consumers, would now rise accordingly; benefits should be determined by the collective value of their expected outage costs.

This value defines the maximum the electric system operator should pay for reserves in the form of load reduction. This result has important policy significance, since under these conditions the valuation of load reduction as reserves bears no necessary relationship to the prices of dispatched energy or generator-supplied reserves, either prior to the emergency or after system reliability has been restored. In other words, market prices cannot be used solely as a guide to setting the price for load curtailments needed during emergency situations to restore system reliability to design levels. Payments for load curtailments reflect their value as a public good used to restore system

security because there is no alternative supply of the good, and consequently they should be based on the expected value of the aggregate losses avoided. [6]

In summary, substantial benefits accrue to all stakeholders when customers participate in wholesale markets as resources under the same market rules as do generation assets. Retailers and customers motivated by the potential private benefits could realize some, and perhaps most in a more mature market setting, of those benefits through RTP-type programs. But, given the low incidence of such programs in either competitive or regulated markets, it will likely be some time in the future, if ever, before such natural price response is actually exercised. This means that retail customers are less likely to realize the potential benefits from restructuring. Alternatively, the full benefit of load curtailments as dispatched reserves can be immediately realized if such actions are taken in concert with prevailing market conditions, which can only be effectively accomplished by the ISO. The criticisms of such programs pale by comparison to the potential gains from giving customers the ability and incentives to effectively participate in wholesale markets. Such opportunities should be explicitly incorporated in the design of wholesale electric markets.

[6] Some have construed this result as meaning that the implicit value of reserves to customers should establish the market price because it reflects the scarcity, which should then properly inure to generation assets. The NYISO allows dispatched DR capacity resources to set the market-clearing price under specific conditions, which in effect transfers some of the gains back to consumers.

1. Introduction

There is near universal agreement that newly created wholesale electricity markets need the discipline and guidance that can only come about when retail customers respond to contemporaneous prices. However, the best way to accomplish this result is the subject of growing debate, the outcome of which has substantial consequences for the role of the Independent System Operator (ISO). Some argue that such customer participation should come through load management, time-of-use, and real-time pricing (RTP) services offered by regulated and competitive retailers; a few proponents would make such services mandatory (e.g., Ruff, 2002 and Borenstein and Holland, 2002).

Some legacy load management programs still exist, but they are valued and operated based on protocols that must be overhauled to reflect the exigencies of wholesale competition. Participation in TOU rates in regulated jurisdictions has been very limited, and it is almost nonexistent in the portfolios of competitive retailers. Finally, RTP programs implemented in the 1990s showed great promise, but they were only adopted by a few utilities that offered service mostly to large commercial and industrial customers. One utility (Niagara Mohawk Power Corporation) implemented a competitive RTP rate as the default service for the largest customers upon inauguration of retail choice, but to date, few others have followed suit. For these reasons, and because price response is so vital to the development of robust wholesale markets, the newly formed ISOs, with prodding from the FERC, have begun to develop and implement programs designed to induce customers to respond to their market prices.

While the approaches differ in degree and method, most would now argue that short-run demand response (DR) programs would facilitate customer participation in wholesale markets and thereby contribute to market efficiency. To improve market efficiency, DR resources must be evaluated on a comparable basis with generation resources for the purpose of RTO scheduling and dispatch operations. If programs are designed properly, DR resources supplement the generation portfolio during emergency situations, and, at other times, compete directly against generation. With proper integration into the ISO's operations, the outcome is equal pay for equal performance, which we demonstrate is the optimal utilization of such resources.

Despite the growing enthusiasm for these new DR programs, there has been limited discussion in the literature to demonstrate exactly how these programs can contribute to market efficiency, the management of market risk, and overall social welfare. This report examines these important issues.[7] The paper is organized in a way that sequentially explores the welfare implications of DR programs for the markets for both the electricity commodity (energy) and for contingency reserves (capacity).

It is well known from economic theory that there are potential welfare gains in moving from average cost pricing to marginal cost pricing (e.g., Spulber, 1989, Borenstein and Holland, 2002 and Stoft, 2002). New competitive wholesale electricity markets seek to achieve these gains by exposing load serving entities (LSE) and other commodity providers to hourly prices, which, in theory at least, should approach marginal costs and could differ dramatically between peak and off-peak periods. Beginning with the status quo (e.g., customers paying conventional fixed-rate tariffs) as the point of comparison, we demonstrate that unless at least some retail customers also see these prices and adjust demand accordingly, much of the anticipated welfare gains from the introduction of competition may be significantly diminished, or worse, not realized at all. A strength of our analysis is that it is not tied to an explicit program design, and our results are valid for a wide range of circumstances. Thus, in discussing the DR programs needed to realize these benefits, we focus on the impacts of DR curtailment, however induced, and that in evaluating the social benefits, if inducement are offered, these must be weighted against the benefits realized. We adopt this strategy to establish the value of demand response. From this foundation, policy makers can consider what, if any, inducements are appropriate to get customers to respond to market prices and generate benefits that inure to all market stakeholders.

To make the analysis more accessible, some arguments are formulated both algebraically and geometrically in the text. In the algebraic formulation in Section 2, we demonstrate differences in quantities demanded when firms face a flat tariff compared to

[7] While the focus of much of this analysis is on price-responsive load management programs, some of the most important results, particularly those related to the welfare implications in the market for the electricity commodity, are similar to those for real time pricing (RTP) programs such as those discussed by Borenstein and Holland (2002). Where appropriate, these similarities, and possible differences are highlighted. However, in the section of the paper that focuses on reserves markets, our approach is distinctly different from that taken by these authors and is designed to answer quite different questions.

when they face separate peak and off-peak prices. In Section 3, we use graphic illustrations to support a discussion of various market situations and delineate measures of both consumer and producer surplus for each case. These situations include a case in which demand can fully adjust to price in a competitive electricity market. We then show that current competitive markets at the wholesale level lead to welfare losses if retail customers continue to face fixed tariffs, thus having no incentive to respond to price signals. Finally, we show that by introducing DR programs in which some customers are paid to reduce load when prices are high, many of these welfare losses from the lack of price-responsiveness can be avoided.[8]

In the case of electricity markets, the realization of welfare gains is further complicated because many customers are unwilling to continually adjust electricity usage, because the transactions costs of doing so outweigh the potential benefits.[9] A small portion of customers may be able to profitably adjust demand to real time prices. For these customers, the 'transactions' costs of demand response may be low because of the nature of how they use electricity, the presence of control technology, or access to on-site generation. These customers would be receptive to services whereby their energy usage was priced at prevailing wholesale market prices, such as those produced by the spot day-ahead market administered by ISOs. If the RTP experience of the past dozen years is an indication, the number of customers that will volunteer for such service is small. There are, however, indications that a third group exists: customers that are willing and able to respond to price, but only on a limited basis.[10] If there are sufficient numbers of such customers, then much of the deadweight welfare loss can still be avoided through

[8] This is in contrast to the situation where a commodity can be stored, in which case the welfare gains come directly through price stabilization (e.g. Just et al., 1982). Welfare gains refer to the net increase in consumer and producer surplus without regard to the distribution of the gains. Further, the amount of DR needed is both an empirical question, as one must characterize the market supply curve, and one with political overtones, as some might argue that displacing generators with DR has deleterious long-term impacts on the supply market (Ruff, 2002).

[9] Clearly, if most customers were able and willing to adjust electricity usage case, standard RTP programs, which by some are considered the economist's first-best solution, could perhaps be implemented on a broad scale. However, RTP programs have seen only limited application, and most of them are not what economists have in mind--customer facing the market price for any and all usage -- but instead involve a hedge, using a two-part model, that seems to affect both their willingness to participate and to adjust electricity use in response to price changes.

[10] Utilities have utilized this principle to operate interruptible, curtailable and load control programs for the past 20 years.

programs that compensate customers for load reductions only when prices are extremely high and/or when system reliability is threatened.

It is difficult to classify customers into these groups because of the paucity of data on observed price responses. Moreover, there may be little reason for customers to self-classify themselves, as proponents of mandatory RTP would suggest, unless they are provided some incentive to do so. This added complication may be further justification for public intervention, particularly in terms of educating customers about how to respond to price or in assisting them in the purchase of technology to facilitate conservation or load shifting, and providing a knowledge base that marketers can use to match service offerings with customer behavior.

While this initial analysis provides a theoretical foundation for policymakers to evaluate the rationale for incentives to promote retail customer price responsiveness, it assumes that system reliability is not directly a factor in setting market prices.[11] To develop a broader set of policy implications, we expand this analysis of the electricity market in Section 4 by accounting explicitly for the supply and demand for both the electricity commodity and reserves. Reserves are additional resources, above what is needed to balance energy usage, that are committed for the purposes of maintaining system reliability at acceptable levels. As such, their value derives from avoiding the consequences of partial curtailments.

Our expanded analysis is presented from the perspective of commercial and industrial customers attempting to maximize expected profits. A similar analysis could also be developed that characterizes how residential customers attempt to maximize expected household welfare. One unique feature of this analysis is that the firm's demand for reserves is treated differently from its demands for other inputs such as labor, capital, or the electricity commodity. In contrast to the standard production function specification, reserves, by reducing the probability of losses due to an outage, are modeled as a damage control agent rather than a conventional input. This damage control agent reduces both the probability of power outages and the firms' economic losses by reducing the difference between planned output and output that would actually be

[11] The analysis focuses on supplying the market's energy needs, which, from a scheduling perspective, includes a built in reserve margin.

realized in the event of an outage. Further, we recognize explicitly that reserves are a public good in the sense that all firms are provided the same level of reserves due to the network character of the transmission and distribution system.[12] This public-good nature of reserves is clearly another rationale for considering what policy actions are warranted to ensure that customers' valuations of service is incorporated into new competitive electricity markets.

In Section 5, we analyze the implications of this specification of energy use, the supply of load reduction reserves, and the value of reserves for five separate situations:
Case 1. The firm maximizes expected profit for a given level of system reserves.
Case 2. The firm is allowed to reduce load below its typical usage level, called the customer baseline load (CBL); this load reduction is recognized by the ISO as an addition to system reserves.
Case 3. The public good nature of the system-wide reserves is considered explicitly.
Case 4. The combined load curtailment rents of firms are assumed to affect electricity price.
Case 5. Generators can supply reserves over and above the fixed level in Case 1; customers can affect prices of reserves and electricity.

In Section 6, we summarize the policy implications of our analysis for both electricity and reserve markets. Drawing from results in Section 5, we discuss various public policies and programs (e.g., "emergency" DR programs) that can internalize and reflect the external benefits associated with the public-good nature of system-wide reserves to customers. In addition to determining the level of payment justified for load reductions, we also discuss the role of augmented investments in education and technologies, perhaps through a system benefits fund, to enable firms to be able to respond to electricity price and financial assistance for purchasing equipment to effect or monitor load shifting behavior.

[12] Network constraints may result in some areas having higher inherent reliability than others, but within those delineations network reliability is the same for all customers. Although not often couched in these terms, it is precisely because system reserves are "public goods" that ISO's are charged with the responsibility of maintaining system reliability. As is seen below, customers, acting only in their own self-interest, would not purchase sufficient reserves to maintain system reliability.

2. Customer Electricity Demand Under Fixed Tariffs vs. Market Prices

To begin this analysis, we assume that a firm j uses four inputs to produce its product denoted as X_i^j, (i = 1,…,4), where X_3 is off-peak electricity and X_4 is peak electricity and the firm's output is denoted as Q^j . The first two inputs could be labor and capital, or other categories of inputs to the firms' production process.[13] The firm's production function F^j (.), is defined as:

(2.1) $Q_j = F^j (X_{1j}, X_{2j}, X_{3j}, X_{4j})$.

It is assumed to have the standard properties, notably concavity in the inputs X_{ij} (Beattie and Taylor, 1985).

We also specify the following definitions for the variables of interest:

P_j = Price of firm j's output;

P_1 = Price of input 1;

P_2 = Price of input 2;

$P_3 = S_3(X_3)$ = Inverse (price dependent) market supply schedule for off-peak electricity;

$P_4 = S_4(X_4)$ = Inverse market supply schedule for peak electricity;

$X_3 = \sum_j X_{3j}$ = Total demand for **off-peak** electricity;

$X_4 = \sum_j X_{4j}$ = Total demand for **peak** electricity; and

T = Fixed tariff for electricity--a weighted average of off-peak and peak prices, weighted by the proportion of total electricity demanded in the two periods.[14]

If electricity suppliers (generators) offer to supply electricity at marginal cost, then the supply relations above, $S_3(X_3)$ and $S_4(X_4)$, are the horizontal summations of the individual electricity suppliers' marginal cost curves. It is convenient to treat the electricity supply side of this analysis as consisting of a single profit maximizing generator whose off-peak and peak cost functions are defined by $C_3(X_3)$ and $C_4(X_4)$, with corresponding marginal cost functions given by $S_3(X_3)$ and $S_4(X_4)$, respectively. Mas-Colell (1995, p.147-48) demonstrates that this is equivalent to the supply behavior of an industry composed of M price-taking firms whose aggregate cost functions are given by

[13] The output of the firm can be a material product or a service. For simplicity if exposition, we will refers to the output as a product.

[14] Borenstein and Holland (2002) also demonstrate that this is the fixed tariff that covers the retail electricity provider's costs.

$C_3(X_3)$ and $C_4(X_4)$, respectively.[15] We also assume that both supply curves yield positive prices and therefore their slopes are positive. Further, it is reasonable to assume that the slope of the supply curve and its price flexibility of supply (the inverse of the price elasticity) is greater in the peak period than it is in the off-peak period, as follows:

(2.2) $S_4'(X_4) >> S_3'(X_3) > 0$.

This would be true for all values of X_j, but it is likely that the individual hourly values of X_4 of interest in the analysis that follows would always be greater than those in X_3.

2.1 Firm Profit Maximization Based on a Fixed Retail Tariff

To maximize firm profits, Π_j, defined by the difference between revenue and costs under a fixed and flat tariff for electricity, the firm demanding electricity is confronted with the following problem:[16]

(2.3) Max. $\Pi_j = \{P_j F^j (X_{1j}, X_{2j}, X_{3j}, X_{4j})\} - P_1 X_{1j} - P_2 X_{2j} - T (X_{3j} + X_{4j})$.

The first term in {} defines the firm's revenue; the second two terms represent other input costs and the last term is sum of off-peak and peak electricity costs at the fixed tariff, T. The first-order conditions for a maximum are defined by:

(2.4) $\partial\Pi_j / \partial X_{1j} = P_j \partial F_j / \partial X_{1j} - P_1 = 0$

(2.5) $\partial\Pi_j / \partial X_{2j} = P_j \partial F_j / \partial X_{2j} - P_2 = 0$

(2.6) $\partial\Pi_j / \partial X_{3j} = P_j \partial F_j / \partial X_{3j} - T = 0$

(2.7) $\partial\Pi_j / \partial X_{4j} = P_j \partial F_j / \partial X_{4j} - T = 0$

where: $P_3 = P_4 = T$.

These are the standard first-order conditions, where the value of the marginal product of each input is equated to its price (Beattie and Taylor, 1985). In this situation, the unit cost of electricity (T) to the firm of both off-peak and peak electricity is the same. We denote the optimal levels of off-peak and peak electricity consumption for firm j, the quantities that satisfy equations 2.6-2.7, by X_{3j}*and X_{4j}*, respectively. Wholesale prices

[15] This is an idealized representation of current electricity markets where competitive firms coexist with regulated firms providing service under administratively determined prices.

[16] Many tariffs include charges for both energy consumed and maximum demand, which leads to a non-linear rate structure, but one that is not implicitly time-differentiated. However, in the short run most customers see the demand charge as inherently dictated by their technology and business schedule, neither of which is readily adjustable. Therefore, the demand charge can be treated as a addition to the energy charge defined by total energy divided by the demand charge.

in the two periods are determined by equating supply in each period with the fixed aggregate demand, which in the short run is determined by the sum of the demands for each firm. These are the quantities of electricity that the firm's supplier (e.g. a load serving entity (LSE)) must purchase from electricity generators (or from the spot market) to serve its customers' demands under the provisions of the fixed tariff.

The corresponding market equilibrium prices are determined by solving the price-dependent supply relations for the aggregate market demands:

(2.8) $P_3^* = S_3(X_3^*)$, where $X_3^* = \sum_j X_{3j}^*$ and

(2.9) $P_4^* = S_4(X_4^*)$, where $X_4^* = \sum_j X_{4j}^*$.

There is no loss of generality to assume that $P_4^* > P_3^*$, the peak price is higher than the off-peak price. It is important to re-emphasize that these electricity demands are the ones revealed by the firms facing the tariff, not those that would result if firms had paid the wholesale prices. Further, for the wholesale supplier to cover its costs, the tariff must be set at the wholesale level to equal a weighted average of the two prices, where the weights are the proportion of total electricity consumed in each period. Accordingly, the equilibrium prices of peak and off-peak electricity that the generator would charge the load serving entity (LSE) would bracket the flat tariff (e.g., $P_4^* > T > P_3^*$).[17]

The social optimum requires that customers in making resource usage decisions equate the value of the resource to the firm with the marginal cost of supply. The inefficiency of this solution in terms of off-peak and peak electricity usage allocation is due to the divergence of the peak and off-peak prices from the fixed tariff. A firm's derived demand for electricity is determined by a downward sloping value of the marginal product schedule (e.g. equations 2.6 and 2.7). Consequently, at the fixed tariff, T, a firm uses electricity during peak periods in excess of the quantity at which the value

[17] Without loss of generality, it is convenient to assume that these prices and the retail customer demand curves are net of a constant wholesale margin. For this reason, we can for the most part refer to generators and retail suppliers of electricity interchangeably. Since the wholesale margin is assumed to be netted out, the remaining revenues collected from customers by retail suppliers are passed on to the generators to purchase the electricity. It is useful at this point to note also that in their discussion of RTP, Borenstein and Holland (2002) view the market from the perspective of a retail electricity provider (often called load serving entities) that purchases electricity from generators to serve end use retail customers. In so doing, they assume that the retail providers know the individual customers' demand curves, as is the supply curve from the generators. As is seen below, we view the market from the customer's perspective to examine the implications for input use and input valuation directly. The retail supplier's margin in netted out for convenience. However, the main results of having customers respond to price remain the same, regardless of the perspective from which they are viewed.

of additional electricity to the firm is equal to the actual cost of supplying electricity, because the tariff rate is below the actual contemporaneous supply cost. Conversely, T is above the marginal cost of supplying electricity during off-peak periods, and as a result the firm uses less than the efficient amount of electricity in off-peak periods. By paying the average price of electricity, the firm's electricity usage in neither the peak nor the off-peak is at its socially optimal level.

As long as the peak and off-peak supply relations have dramatically different slopes,[18] if customers instead faced prices that equal supply cost during the two periods, aggregate demand would fall during the peak period, as would the peak wholesale price, supply held constant. Similarly, demand would increase during off-peak, and off-peak prices would rise accordingly. These effects are demonstrated in the analysis of time-of-use rate below.

2.2 Firm Profit Maximization Based on Wholesale Peak and Off-Peak Prices

One way to expose customers to underlying wholesale price variability is to offer a retail rate that separates the hours of the year into those periods characterized by high prices (the peak period) and those with low prices (the off-peak period). Such a time-of-use (TOU) rate forces customers to evaluate their usage patterns, which were optimized for a flat rate, relative to price differences, and ascertain if they can benefit from switching usage from the high price to the low price period. To maximize profits when facing differential peak and off-peak prices, the firm now solves the following problem:

(2.10) Max. $\Pi_j = P_j F^j (X_{1j}, X_{2j}, X_{3j}, X_{4j}) - P_1 X_{1j} - P_1 X_{2j} - P_3^c X_{3j} - P_4^c X_{4j}$.

In this case, the firm faces time-differentiated, competitive wholesale electricity prices, which are determined by the following:

(2.11) $P_3^c = S_3(X_3^c)$, where $X_3^c = \sum_j X_{3j}^c$ and

(2.12) $P_4^c = S_4(X_4^c)$, where $X_4^c = \sum_j X_{4j}^c$.

The superscript c denotes the case where peak and off-peak rates are different. Assuming an interior solution exists, the necessary and sufficient conditions for an

[18] There has been much discussion of the existence of the so-called hockey-stick shape of wholesale supply curves and their implications for price determination (Caves et al., 2000). Boisvert et al., (2002) document empirically that the slopes of these for peak and off-peak periods are dramatically different in the New York electricity market.

equilibrium are that: a) the market clears, i.e. aggregate demand for electricity equals supply; b) the prices received by generators are equal to the marginal cost of supply, and c) the value of the electricity to any firm is equal to its price (Mas-Colell, 1995). In other words, a single (but different) price prevails in each of the peak and off-peak periods, and all suppliers and customers adjust their behavior accordingly.

Focusing on a firm's electricity use, it solves the profit maximization problem defined by (2.10) and the first-order conditions necessary and sufficient conditions are now as follows:

(2.13) $\partial \Pi_j / \partial X_{1j} = P_j \partial F_j / \partial X_{1j} - P_1 = 0$

(2.14) $\partial \Pi_j / \partial X_{2j} = P_j \partial F_j / \partial X_{2j} - P_2 = 0$

(2.15) $\partial \Pi_j / \partial X_{3j} = P_j \partial F_j / \partial X_{3j} - P_3^c = 0$

(2.16) $\partial \Pi_j / \partial X_{4j} = P_j \partial F_j / \partial X_{4j} - P_4^c = 0.$

Since the derived input demand schedule for the firm operating under a well-behaved production function is downward sloping, it is clear that the new competitive quantities demanded of off-peak and peak electricity are $X_{3j}^c > X_{3j}^*$ and $X_{4j}^c < X_{4j}^*$, respectively.[19] The superscript ^{c} is used to distinguish the competitive equilibrium prices and quantities when firms face time-differentiated rates from those that would result when firm's face fixed tariffs, which are denoted by a superscript asterisk (*).

One can view the adjustment from when the firm faced a flat rate tariff to this new equilibrium as occurring in steps. In the first step, firms are confronted with the wholesale prices implied by the quantities demanded (X_{3j}^*, X_{4j}^*) under the flat tariff. As seen above, and using the asterisk notation, these prices would be P_3^* and P_4^*. Firms will recognize under this price schedule, that the marginal value of the last unit of demand on-peak (off-peak) is less (greater) than P_4^* (P_3^*), and reduce (increase) demand

[19] A firm's derived demand curve for electricity can be established by first solving these first-order conditions for the profit maximizing levels of the X_{ij}'s. These input levels are in turn substituted into equation (2.10). The derivative of the resulting indirect profit function with respect to an input yields the derived demand for the input (Mas-Colell, 1995).
Borenstein and Holland (2002) assume that it is these demand curves that are know to the retail electricity suppliers. Firms on RTP are assumed not to differ from those on flat rates, but the demand do, since only RTP customers adjust demand in response to price. Aggregate demand in their case can then be characterized by the weighted average of the two demand curves, weighted by the proportions of firms on RTP and on the flat rate. By assuming that all customers either face the fixed tariff or respond to price, our results are comparable to theirs where the proportion of customers on RTP is zero or unity. For proportions in the middle range, the results, in particular the magnitude of the inefficiencies and social welfare losses due to the fixed tariff, would differ only by degree.

on-peak (off-peak) so that $X_{4j}^c < X_{4j}^*$ and $X_{3j}^c > X_{3j}^*$. Further, although no individual firm's actions will necessarily affect prices, the combined changes in demand will do so. Wholesale suppliers will adjust their load bids in the wholesale market accordingly and the new market equilibrium prices will evolve such that $P_3^c > P_3^*$ and $P_4^c < P_4^*$, where conditions (2.15) and (2.16) are met: the competitive off-peak price is higher and the peak price is lower.

Moreover, the new quantity weighted average price to the firm (T^c) will fall (e.g. $T^c < T$). This follows directly from the assumption that the LSE sets the fixed tariff to cover the average cost of electricity purchased at wholesale prices of P_3^* and P_4^*. Since the competitive peak price falls relative to what it was before (P_4^*), and the off-peak price rises relative to P_3^*, a sufficient condition for average price to fall for T to T^c is for peak usage to fall (e.g., $X_{4j}^c < X_{4j}^*$).

In this process by which firms respond to the imposition of marginal prices, the market inefficiencies will disappear. In addition, there will be transfers from generators to retail customers, particularly as the peak wholesale price falls. The nature of these transfers, as well as the elimination of dead-weight losses, another consequence of time-differentiated pricing, are perhaps best revealed through a simple set of diagrams. In the next section, we analyze the load reduction during peak periods that are needed to restore market efficiency, and determine what society can afford to pay to see that these adjustments are realized.

3. A Diagrammatic Welfare Analysis of Competitive Electricity Markets

In this section, we describe a welfare analysis of electricity markets under situations where all firms can adjust their usage in response to price signals, and firms face fixed tariffs with peak and off-peak prices, using geometric diagrams. To the extent possible, we keep the notation consistent with that used in section 2.

3.1 Competitive Electricity Market with Full Capacity to Adjust to Price Signals

Consistent with the model described in Section 2, we assume that the market for electricity is divided into two distinct periods, an off- peak period and a peak period. In this market, generators submit offers to sell un-contracted for capacity and energy to a

last price auction and demand is uncertain. The market price and the amount each generator is to supply are determined simultaneously. These conditions characterize day-ahead wholesale electricity markets such as that run by the New York ISO and PJM Interconnection (PJM), and are also consistent with the standard market design as currently proposed by FERC. They also in principle apply to real-time pricing programs operated by vertically integrated utilities such as Duke Power, Public Service of Oklahoma, and Georgia Power.

We initially assume that customers can make *full* and *costless* adjustments to demand in response to price changes according to established derived demand schedules for electricity that represent the value of the marginal product of electricity to the firm. The situation is depicted in Figure 1. For analytical purposes below, we will consider the peak and off-peak periods separately.

3.1.1 Off-Peak Demand

Figure 1 depicts off-peak ($D_{o)}$ and peak ($D_{p)}$ demand curves for the firm imposed on a single supply curve. The different price/quantity pairs mapped out represent different market circumstances, as described below.

*According to Figure 1, the competitive equilibrium in the off-peak period is at point **Y**. Here, retail customers during off-peak periods follow demand curve depicted as Do in the figure and buy X_3^c at price P_3^c at a total cost of $X_3^c P_3^c$. If the demand curve is net of a constant wholesale margin, M, the retail prices and quantities are the same as in Section 2. The generators supply X_3^c according to supply curve S and are paid P_3^c yielding revenue equal to $P_3^c X_3^c$. Under these conditions, welfare is measured by the sum of consumer and producer surplus:*

- Consumer surplus is the area under the demand curve Do and above the price line P_3^c, indicated in Figure 1 by the box labeled i and the triangles h and r.

- Producer surplus is the area in Figure 1 above the supply curve S and below the price line P_3^c, as indicated by (j + k + n).
- Welfare is the sum of the producer and consumer surpluses, area defined by {h + i + r} + {j + k + n}.

3.1.2 Peak Demand

The competitive equilibrium for the peak period if customers respond to price changes is indicated by Z'' in Figure 1, the intersection of the peak demand curve D_p and price P_4^c (see Fig. 1). During periods of peak demand, retail customers buy X_4^c at a price of P_4^c and a cost of X_4^c P_4^c, where the demand curve is net of a constant wholesale margin, M, similar to the case for the off-peak period. The generators supply X_4^c and are paid P_4^c, and they receive revenues of P_4^c X_4^c. The measure of welfare is again given by the sum of consumer and producer surplus.

- Consumer Surplus is the area to left of and above D_p and above P_4^c , the area (a + b).
- Producer Surplus is the area above S, to the left of Dp and below P_4^c , the area (h + i + r + j + k + n + s' + g).
- Welfare is the area {a + b} + {h + i + r + j + k + n + s' + g}.

If this were a market for a storable commodity whose production takes place prior to knowing demand conditions, one could develop a buffer stock scheme to even out supply and demand. That is, the buffer stock agency buys when supply exceeds demand (off-peak) and sells when demand exceeds supply (supply). Under these conditions, Just et al. (1982) show that society gains from such price stabilization actions if price is set at the weighted average of P_3^c and P_4^c, with the weights being the probability of each state (Just et al.,1982).

Unfortunately, electricity is not storable, so the analysis of Just et al. (1982) does not apply directly to these circumstances. Further, under current retail market conditions most retail customers can still buy electricity at fixed rates, but their suppliers face fluctuating market prices.[20] To see the value of inducing price responsiveness, we must compare the case just illustrated, where demand can fully respond to price, with the previous situation whereby retail customers can use any amount of electricity at fixed prices.

[20] For many customers, it is not practical to adjust demand in response to price changes; the transactions costs (outage costs plus costs of administration, meters, etc.) of doing so are very high. This means that the two aggregate demand curves in Figure 1 are the horizontal sum of many individual demand curves, most of them completely inelastic (e.g. completely vertical), or nearly so.

3.2 Competitive Wholesale Electricity Market with Retail Demand Served at Fixed Prices

To complete this part of our illustrative welfare comparison, we again look at the off-peak and peak periods separately, at least initially.

3.2.1 Off-Peak

We begin by examining the outcome for the off-peak period under the flat tariff T, again assuming that demand curves are net of any wholesale margin. For ease of discussion, we introduce Figure 2, which isolates the off-peak period situation.

In off-peak periods, the fixed tariff (T in Figure 2) is set above the off-peak market price, because as discussed in section 2, for the wholesaler to cover the cost of both peak and off-peak power purchases, T must be a weighted average of the peak and off-peak prices.[21] The equilibrium in this case for the customer is indicated in Figure 2 by point X, with the firm consuming quantity X_3^*. At this point, as illustrated in Figure 2:

- Consumer Surplus = (h)
- Producer Surplus = (i + j + k) (i + j go to the customer's load-serving entity (LSE); k goes to the generator)
- Social Welfare = {h} + {i + j + k}
- Social loss compared with the competitive market situation where customers can respond to price is: { r (foregone consumer surplus) + n (foregone producer surplus)}.

Compared with the situation described in section 3.1 where customers can respond to price, social welfare is reduced under the flat tariff by the areas r + n, which is called deadweight loss, while consumer surplus, area i, is transferred from customers to the LSE. Transfers do not affect the level of net social welfare, only how it is shared among consumers, generators, and retail suppliers. In this case, the LSE 'stores' the transfer revenues to cover revenue shortfalls in the peak, when its tariff is below its supply costs.

[21] As above, T is a weighted average price, where the weights are the proportion of electricity consumed in each period.

To summarize, social welfare can be increased by offering to sell additional load at the lower price P_3^c. Demand and supply will continue to adjust, until the equilibrium point Y is reached. At Y:

- Producer surplus increases by an amount equal to the area n
- Consumer surplus increases by an amount equal to the area r, which either the supplier retains unless it lowers the price of all X_3^c to the customer, in which case the customer would realize the full benefit, and area i is transferred back to consumers.

Regardless of who retains the increase in producer and consumer surplus, Y is preferred socially to X since it represents the point at which MC=MR, resulting in the socially optimal use of resources.

3.2.2 Peak Period

We next examine the situation in the peak period in a similar fashion, using Figure 3 for ease of exposition. When customers are faced with a fixed tariff, the equilibrium point will be at point Z in Figure 3, where the retail price is fixed at T and quantity consumed is X_4*. The flat tariff also leads to inefficiencies in the peak period because for demand greater than X_4^c, the usage price, which represents value to the firm given by points on the demand curve, is below marginal cost (e.g. the supply curve). The use of electricity whose value in production is below the cost of electricity results in deadweight loss in welfare to society represented by the combined area d + d′.

The distribution of producer and consumer surplus in the peak period case requires care to disentangle. We know that on average the price T covers the cost of the LSE purchases of energy to serve the customers both during peak and off-peak periods. Therefore, in looking at Figure 3, we can assume that expenditures by LSE to buy power at peak prices above T is effectively collected from the customer through off-peak sales at T which is above the supply cost, as discussed above. If the supply curve were indeed flat, as it effectively is from the customer's perspective when facing a fixed price of T, consumer surplus at price T (Figure 3) would be: a + b + g′ + f + e, and there would be

no producer surplus. The wholesale suppliers and in turn generators would be paid T for each unit, and that payment would equal marginal cost.

However, implicit in the fixed tariff T (determined simultaneously with X_4*and X_3*) is a payment of X_4*[P_4*- T] (and quantity weighted) to cover the wholesaler's cost of X_4* over and above T. This amount is transferred by the LSE to the generator and is equal to the combined area b + c + d + d′ + g′ + f + e. The areas b + c + f + e are consumer surplus transfers from the customer to the generator (through the LSE levelizing tariff revenues) during the peak period and thus augment producer surplus above the level s′. The final result is that consumer surplus = a, and producer surplus = s′ + b +g' + c + d + d'. The generator also receives payments (economic rents) equal to the combined area d′ + d, which represents additional costs to the customer resulting from the inefficiency in pricing all usage at T rather than at the true differential prices that reflect the marginal cost of supplying electricity. From society's perspective, the additional resources needed to produce X_4* – X_4^c (e.g., consumption over and above the optimal level) would have been better allocated to other uses; thus the combined area d′ + d is lost to detriment of society, and is referred to as the deadweight loss.

The challenge facing electricity market designers and policy makers is how to design retail programs that can reduce or eliminate altogether the size of these deadweight losses. There is perhaps no single solution to the problem, but we can highlight the important issues by illustrating the impact of a Demand Response (DR) program, which encourages customers to bid P_4^c to provide load reduction in the amount [X_4*– X_4^c], thereby eliminating the deadweight loss. Payments to those that accomplish this load reduction would be the combined area s″ + e + d′ (see Figure 3). *As long as this area is less than the deadweight loss of d¢ + d, then social welfare is unequivocally improved.* In other words, for there to be an increase in net social welfare for a DR program, (s" + e) < d; these areas are illustrated in Figure 3.[22]

[22] Borenstein and Holland (2002) provide an analysis of the second-best optimum if customers are to remain on flat tariffs. Their arguments are summarized here because through further analysis, one may be able to discover an algebraic relationship between these areas, although such an analysis is not done in this paper. As stated above, Borenstein and Holland (2002) shows that the quantity weighted average price, T, is the flat tariff that will cover the costs of retail electricity suppliers. However, this is not the flat tariff that provides the second-best welfare solution if retail customers stay on flat tariffs. Instead, they show that the flat rate tariff that minimizes the dead weight loss is one in which the price weights are the relative slopes

The size of these two areas is clearly an empirical question.[23] From a policy perspective, we can view this situation in two different ways. The first relates to the characteristics of supply and demand if firms have an incentive to respond to price and achieve the equilibrium defined by point Z'' in Figure 3. Viewed from this perspective, it is clear that as the supply curve becomes steeper (e.g. pivoting counter clockwise around point Z''), the net welfare from a DR program increases because the area d becomes larger. Similarly, if the initial demand curve were less price responsive (made steeper by pivoting clockwise about the competitive equilibrium Z'') the net welfare calculation would also move in favor of the DR load, as the areas e and s'' would both become smaller. In summary, the potential welfare gains from DR load programs are highest in situations where both the supply and demand curves are initially extremely price inelastic ("steeper"). These are the very circumstances that have lead to price spikes that disrupt newly formed wholesale markets, and militate for the implementation of programs that can mitigate the circumstances that result in welfare losses.

Consequently, from a societal perspective, it makes sense to promote programs that expose customers that will respond to market prices during the peak period when they are high. This view provides a basis for understanding the potential gains from implementing DR programs. Prior to program implementation, firms would be facing a fixed tariff and consuming at point Z in Figure 3. Thus, if we take this as a starting point, the welfare gains from a DR program can be increased if: a) firms can be encouraged to reduce overall peak demand (e.g. resulting in a shift in D_p to the left) and/or, b) if the supply curve is sufficiently steep, firms can be encouraged to be more price responsive just during peak periods (e.g., resulting in D_p pivoting counterclockwise around point Z).

---

of the peak and off-peak demand curves. This rate may be higher or lower than the value of T. This is an important result, but it depends on the supply curve being perfectly elastic up to system capacity, and vertical at that point. If supply elasticities are in between these extremes, the second-best fixed tariff would also likely involve the slopes of the supply curves as well, although this is not derived explicitly here. At some time it would be useful to derive this more general result, although it is not critical to the validity of their argument.

As Borenstein and Holland (2002) also point out, one difficulty with this second-best fixed tariff does not necessarily allow retail suppliers to cover their costs. However, these costs can be covered along with achieving the second-best solution under competition through a tax or subsidy that is the quantity weighted average of the new second-best flat tariff.

[23] For convenience, Figure 3 was drawn assuming linear supply and demand curves, but this representation may in fact distort the size of the areas being compared.

The former situation calls for permanent changes in consumption patterns by introducing time-of-use pricing. The latter situation is more effectively accomplished by exposing customers to prices (or incentives derived there from) when such market conditions occur and the societal benefits justify the incentive required to induce the desired behavior.

4. Modeling Firms' Demand for Electricity and System Reserves

The diagrammatic analysis in Section 3 addresses the social welfare losses inherent in markets for electricity when the supply is upward sloping, but demand is nearly vertical. Retail customers still purchase load at fixed tariffs equal to the average wholesale price (plus an appropriate wholesale margin). Under such pricing, customers have no incentive to adjust load down (up) when wholesale prices are high (low) until it is too late: when the price impacts have become incorporated into the subsequent fixed tariff, or retroactively applied to the past month's consumption.[24] The result is deadweight losses in the electricity commodity market in both peak and off-peak periods. If, however, some customers are price responsive, the analysis in Section 3 demonstrated that much of this deadweight loss could be mitigated through DR programs where customers are paid to curtail during high-priced periods. Such a system improves social welfare.

The analysis in Section 3 ignored the fact that customers' willingness to curtail load during peak periods not only affects peak prices, but also constitutes a resource that can contribute to system wide reserves. To characterize the full extent of the welfare implications of price responsiveness in new electricity markets, we extend our analysis to consider the role of reserves explicitly. Because of the additional complexity, it is now necessary to formulate the analysis mathematically, but some simplified diagrams are introduced to better illustrate the essential points.

[24] Many regulatory rates reflect market price volatility by adjusting rates based on the most recent month's experience. This ensures the supplier that it recovers costs, but the adjustment is too late to abate the price volatility from which it is derived, and might led to a delayed reaction reduction in usage during periods when prices are low, thus incurring additional deadweight losses.

4.1 The Firm's Production Function

The purpose of this analysis is to portray how a firm's demand for electricity as an input to a productive process is affected by different market conditions and assumptions regarding firm demand for reliability. It is therefore tempting to conduct the analysis in terms of a dual, indirect profit or cost function for the firm so that factor demands could be represented directly as functions of input and output prices. However, that specification is difficult to estimate from the data utilities typically have on their customers. We instead work with the primal formulation, because doing so underscores the implications of firm input substitutability for improving electric system performance and illustrates the effects of adopting energy efficient or load shifting technology. The value of this approach becomes quite apparent as the discussion proceeds.

In order to differentiate the role of electricity commodity and reserves for the individual customer, it is essential that we begin by specifying a general production function for a representative firm.[25] However, the value of electric system reserves to the firm's production process is unlike that of conventional inputs, such as labor and materials. The level of reserves determine how often firms experience an outage and as a result incur additional production cost or suffer losses, and thereby establishes how effective the other inputs are relative to their full potential. If the firm experiences no electric service outages, then inputs achieve their physical productive potential. Outages however disrupt the firm's ability to achieve that end, as production must be rescheduled or in some cases foregone altogether.

Thus system reserves serve as a damage control function. In this regard, we adopt an analysis similar to that of Lichtenberg and Zilberman (1986) who imbed a production function in a model to represent a firm's decision to control pest damage to agricultural production through the application of pesticides.[26] The level of application of pesticides by the firm determines pest damage and the ultimate productive capability of seeds, fertilizer, labor, etc, but it also effects production costs and therefore profits. In the case

[25] As before, we characterize the decision process of a commercial or industrial customer through this production function, assuming the firm is maximizing profits. A corresponding analysis can be constructed for residential customers assumed to maximize utility subject to a budget constraint.

[26] In that paper, they argue that such a model is applicable to pest management efforts: 1) to reduce crop losses by reducing the size of pest populations at certain times of the year; 2) to vaccination programs to reduce the susceptibility to infection; or 3) to the installation of sprinkler systems etc. to reduce the damage from fire.

of electricity demand, it is system-wide reserves that act as the control agent. Through increases in reserves, losses due to power outages are avoided and/or the probability of an outage is reduced, thereby increasing the value of the other productive inputs used by the firm.

The level of the system reserves is of concern to the firm, but an increase in reserves does not increase the production of the firm directly, as is the case of employing additional standard types of inputs such as land, labor, capital, and energy. To understand the contribution of electric system reserves to a firm's production, it is instructive to view the actual output of a firm in terms of two separate components: a) potential output, the maximum quantity of output attainable from the application of a specific quantity of inputs, and b) the losses in that productive potential due to system-wide outages. The productivity of the damage control agent, in this case system-wide reserves, must therefore be defined in terms of its contribution to damage (outage) abatement. It is clear that damage, and thus the value of the abatement, is limited both by potential output and by the total losses due to the damaging agent—in this case the outage. However, total damage to the firm can be no larger than the value of potential output and no smaller than zero.

Viewed in this way, the appropriate restrictions on the effects of abatement (system reserves deployed to reduce the instances of outages) can be captured by specifying an abatement function G*(R|N) defined as the proportion of the destructive capacity of the outage eliminated by the application of a specific level of the system-wide reserves, R, for a specific system state N.[27] For a given system state, N, system reserves is the major argument in the abatement function; the function G* is equal to: [1 - the loss of load probability (LOLP)]. Thus, the function G* will conveniently have the properties of a cumulative probability distribution defined on the interval (0,1). When G* = 1, there is the complete abatement of the destructive capacity since the probability of an outage is

[27] It is important to note that G is likely to be a function of variables other than the level of reserves, such as system states, the availability of imported power, unplanned generator or transmission outages. One could think of these factors as being captured by N and shifting the function G* so as to change proportion of the destructive capacity of an outage for any given level of reserves. Since in this model we do not distinguish between peak and off-peak periods, the differences in the G* function at the two different times could be captured in N. In developing a more specific empirical formulation in section 5, we do account for both peak and off-peak electricity use. At that point, we essentially return to the input specification in equation (2.1).

zero. At the other extreme, when G* = 0, destructive capacity is at its maximum because the probability of an outage would effectively be equal to unity. The function G* is monotonically increasing in reserves, and will approach the value of one (1) as reserves increase in the limit to infinity. The derivative of G* with respect to R represents the marginal increase in system security (marginal decrease in the probability of an outage) as the level of reserves, R, increases; analytically, it is simply the density of G*(R|N).

Using the notation from Section 2, the actual output production function for firm j, Q_j, can now be written as a function of the direct productive inputs and the outage abatement function, G* (R|N):[28]

(4.1) $F^j (X_{1j}, X_{2j}, X_{3j}, X_{4j}, G^* (R|N))$.

The production function, F^j (.), is assumed to have the standard properties, notably concavity in Z_j, E_j, and G* (Beattie and Taylor, 1985).

While it would be possible to develop this analysis using this general form of the damage control function, it is instructive to assume that damage control can be specified as two distinct elements. We first characterize the probability of an outage as [1 – G^j(R|N)], where, as above, G^j (R|N) is the probability that no outage occurs for firm j.[29] This probability function depends on system-wide reserves, R, and is conditioned by the system-state, N. Second, we specify a damage function D^j (R|N), which is also a function of R and N; it is defined as the proportion of output that is lost during an outage.[30] There

[28] In their attempt to understand the properties of this production specification, Lichtenberg and Zilberman (1986) derive the expressions for the elasticity of demand for abatement (G*) as well as the elasticity of demand for the control agent. In evaluating these expressions, they conclude that because g*(R) has the properties of a density function, then the marginal effectiveness of abatement is always elastic, but that the for any commonly used distributions, the demand for the damage control inputs (reserves in this case) would be everywhere inelastic. Further, in general, more elastic the demand for abatement, the more elastic will be the demand for reserves, the control agent.

[29] A similar two-step process for agricultural pesticides is discussed by Fox and Weersink (1995).

[30] Recall from above that the function G will conveniently have the properties of a cumulative probability distribution defined on the interval (0,1). When G = 1, there is the complete elimination of the destructive capacity (the probability of an outage would be zero in this case). At the other extreme, when G = 0, destructive capacity would be at its maximum (the probability of an outage would approach unity). The function G will also be monotonically increasing in reserves, and will approach 1 as reserves increase in the limit to infinity. The derivative of G with respect to R represents the marginal increase in system security (marginal decrease in the probability of an outage) as R increases. Analytically it is simply the density of G(R).

Similarly, the function D will also have the properties of a cumulative probability distribution defined on the interval (0,1). However, when D = 1, the loss of output from an outage is complete, while at the other extreme, when D = 0, the there would be no loss of output from an electrical system outage. The function D will be monotonically decreasing in reserves, and it will approach unity as R increases in the limit. The

are probably many reasons to specify D as a function of R, but the most compelling is that R affects the length of an outage.[31] It is assumed that D is a function of the system-state, N, and therefore it is reasonable to assume that damage would increase with the duration of the outage, eventually at a decreasing rate. While D is specified as a separate function of R and N, it is important to emphasize that it is conditional on the occurrence of an outage. That is, if no outage occurs, there are no output losses.

These two functions are illustrated in Figures 4 and 5 for normal (N_N) and severe (N_S) reserve availability states. In Figure 4, the level of reserves is measured on the horizontal axis and the probability of there being no outage on the vertical axis, where the probability of an outage is measured by unity at the origin and the probability of an outage decreases as the level of reserves increase, depicted by moving from left to right in the diagram. The functions G ($R|N_N$)) and G ($R|Ns$)) in Figure 4 emanate from the origin and approach unity (e.g. a zero probability of an outage) as reserves (R) increase. The probability of no outage for a specified level of reserves, represented by R^o in Figure 3, under a normal system state (G ($R^o|N_N$)) is illustrated by point B in Figure 4, while the probability of an outage is the distance AB. If the system state deteriorates due to an unforeseen transmission or generation problems, illustrated in Figure 4 by s shift in the G function to the right (to the dotted curve $G(R/N_s)$), and the probability of no outage decreases at every level of reserves for example by the distance BC (which is less than distance AB) at a reserve level R^o. Further, as the level of system reserves increases by ΔR, as illustrated by the shift from R^o to R* in Figure 4, we move along the curve for a normal system state (G ($R|N_N$)) from point B to point E, and the probability of an outage declines (i.e., system reliability increases) by a distance ΔG. Put differently, $\Delta G/\Delta R > 0$; improved reliability abates potential damages to the production processes and thereby improves the firm's expected profits, but it does so at decreasing rate.

The relationship between system electricity reserves and system states, and the proportionate losses of a firm, is illustrated in Figure 5. Reserves are indicated on the horizontal axis, while the proportion of the firm's output lost during an outage is

---

derivative of D with respect to R is simply the density of D (R), and it represents the marginal decrease in the proportion of output lost due to an outage as R increases.

[31] This is consistent with the notion that reserves are used to restore the system when an outage occurs. Thus, as reserves increase, the time to restore the system decreases.

measured on the vertical axis. The proportion of output lost during an outage when reserves are at R^o as illustrated in Figure 5, where the state of the system is normal at point C on the normal curve (G ($R^o|N_N$)). Lost output would increase by the distance CB if the damage function D shifts up and to the right to D(R|Ns)), indicating a deterioration in the state of the system. For simplicity, we have the proportion of output lost approaching unity (i.e., 100%) as reserves are reduced to zero. This is an upper bound on losses, but the proportion of a firm's output lost during even an extended outage may never reach this level. As the level of system reserves increases by ΔR (e.g. from R_o to R*), under a normal system state, we move along the curve (G ($R^o|N_N$)), and the proportion of output lost due to an outage declines by a distance ΔD. That is, $\Delta D/\Delta R < 0$.

In summary, under this formulation the expected proportion of firm's output lost due to system outages is determined by [1 – G(R|N)] [D(R|N)]. This can be an extremely small number if the probability of an outage is small, if the proportion of output lost (the firm's outage cost) is small, or if both are small. Alternatively, expected lost output approaches unity (i.e., 100%) in extreme system conditions when outages are very likely. As demonstrated in the models below, the magnitude of these factors helps to explain a firm's willingness to participate in DR programs, and it clearly affects the size of incentives needed to encourage participation.[32] If the above expression is the expected proportion of output lost due to outages, the expected proportion of output remaining during an outage is [1 – G(R|N)] [1 - D(R|N)]. It is this latter expression that is most convenient for use in the models below.

4.2 Profit Maximizing Behavior

Given this specification of the damage control function, we can specify the firm's decision problem as one of maximizing expected profits, as follows:

$$(4.2)\ \max.\ E\Pi_j = [G^j(R^*|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\} + [1-G^j(R^*|N)]\ [1- D^j(R^*|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$$

[32] In response to the elevated likelihood of rotating outages during the California electricity crisis, California utilities offered a program, called the Optional Binding Mandatory Curtailment, which exempted participating customers from outages if they agreed to reduce their usage by 15% whenever reserves margins fell below 5%. This is clearly an appeal to the same behavior that we characterize here; part of a loaf is proportionally better than none.

$$- P_1 X_{1j} - P_2 X_{2j} - T (X_{3j} + X_{4j}),$$

where an individual firm takes the level of reserves and the price of the electricity commodity as given. Note that in this specification the firm faces fixed tariff prices T. This formulation represents the base case, which we will refer to as Case 1, in which the firm buys electricity at a fixed tariff. The solution to this problem can be thought of as identifying a firm's normal usage, which we call the CBL, will serve as basis of comparison against which we measure a firm's compliance with load curtailment opportunities under a DR program and the change in benefits that result.

To demonstrate the implications of this production specification for the firm's profit maximizing demand for electricity and reserves, we consider four additional cases. In Case 2 we assume that the firm can curtail load and that such load curtailments contribute to system-wide reserves. In this case, the firm receives no direct payment for its load curtailment, and therefore it considers only the internalized benefit from an expansion of reserves in making the decision on the level of reserves it will self-supply through load curtailments.

Since customers are connected to a single transmission network, reserves are a public good in the sense that all customers enjoy the benefits from a given level of reserves. Reserves meet the definition of a public good in that the availability of system-wide reserves to one firm does not preclude their availability to other firms (e.g. Mas-Colell et al., 1995; Baumol and Oates, 1988). To examine the implications of the public good nature of reserves, we introduce Case 3 to examine the situation in which the joint profits of all firms are maximized given that the demand for electricity by each firm must not exceed the CBL from Case 1. Again, the firm is not paid for curtailments, but the value of reserves both to the firm and to other firms is revealed. Through this analysis, we can determine what society (or other firms that stand to gain) should be willing to pay for curtailments to internalize the benefits of the public good nature of reserves and compare that with the private value assessed to such curtailments by the firm based only on the value it internalizes.[33] We expect that if reserves are a public good, there will be a gap between those valuations.

[33] Socialization of incentive payments for customers that curtail are recovered by ISO by taxing uses, which is proper because is the other customers that gain from the increased reliability. In this case, society can be

In Case 4 we relax the assumption that firms can have no effect on price when they undertake load curtailments. We examine this case to establish that a primary benefit of load curtailment is to release some generation that effectively augments the otherwise fixed compliment of reserves. However, curtailments can also reduce the demand for the electricity commodity. Thus, when the market is characterized by a nearly vertical supply curve, the load curtailments may also exert substantial downward pressure on the price of electricity. The final case, Case 5, analyzes situation in which both generators and customer loads (i.e., firms) can supply reserves and their actions affect prices in the reserve market.

5. Welfare Analysis of Electricity and System Reserves: Case Study Results

The systematic and serial evaluation of the cases described in Section 4 exposes the character of reserves as a public good, and demonstrates why programs that pay customers to curtail when prices are high are justified and desirable in the absence of other incentives to induce customers to respond accordingly.

5.1 Case 1: The Firm's Maximization Problem and the Demand for Electricity

Based on the imbedded damage function specification, we portray the firm as maximizing expected net return or profit.[34] We adapt the input specification of the firm's production function that was introduced in equation (2.1) in Section 2 and define other variables and relations as needed to incorporate a firm's consideration of the value of electric system reserves. For simplicity, we also assume that only output is lost to the firm due to an outage. More complex models might also account for inventory losses, some savings in input costs from a sustained outage if workers are sent home, expenses on damage mitigation, and health and safety ramifications.[35].

---

though of as the community of electricity users whose reliability is entrusted to the central reliability agent, the ISO.

[34] One could also assume that the firm is risk averse and maximize the expected utility of a concave utility function (e.g. Boisvert et al., 1997), but that would unnecessarily complicate the analysis and not lead to any new insights.

[35] We could accommodate input cost savings simply by distinguishing between other inputs that are used during peak and off-peak periods, but this would only serve to expand the number of first-order conditions unnecessarily. The results would be more complex, but their general nature of the results would remain unchanged.

We begin by assuming that the firm is faced with a fixed tariff for electricity and there is level of system reserves is exogenous to the firm's specific expected outage costs; it is determined by the local reliability council. The profit maximization problem is given by:

(5.1) max. $E\Pi_j = [G^j(R^*|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$+ [1–G^j(R^*|N)] [1- D^j(R^*|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\} – [P_1 X_{1j}- P_2 X_{2j} - T (X_{3j}+ X_{4j})]$.

The first order-conditions for this problem provide the basis for demonstrating the value of reserves to the firm: they are as follows:

(5.2) $\partial E\Pi_j / \partial X_{1j} = \{P_j \partial F_j / \partial X_{1j}\}\{[G^j] + [1–G^j] [1- D^j(R^*|N)]\} - P_1 = 0$

(5.3) $\partial E\Pi_j / \partial X_{2j} = \{P_j \partial F_j / \partial X_{2j}\}\{[Gj] + [1–G^j] [1- D^j(R^*|N)]\} - P_2 = 0$

(5.4) $\partial E\Pi_j / \partial X_{3j} = \{P_j \partial F_j / \partial X_{3j}\}\{[G^j)] + [1–G^j] [1- D^j]\} - T = 0$

(5.5) $\partial E\Pi_j / \partial X_{4j} = \{P_j \partial F_j / \partial X_{4j}\}\{[G^j] + [1–G^j] [1- D^j]\} - T = 0.$

The value of the marginal product of each resource differs from the previous specification in each is multiplied by the expression $\{[G^j] + [1–G^j] [1- D^j]\}$. The value of an additional input is diminished by the probability of an outage times the proportion of output lost. Thus, the marginal value product of each resource is equated to its price divided by this expression, which is less than unity, since G^j and $[1 – G^j]$ both take on values of between zero and one, as does the term $[1 – D^j]$. This is particularly significant in that if firms explicitly account for the possibility of an outage and the attendant losses, they will employ fewer resources and produce less output than otherwise (e. g. than in the problem given by equation (2.3)).

Put differently, since the marginal products of all resources are declining, if we let the expected profit-maximizing levels of the inputs be denoted by a superscript #, then compared with the problem in Section 2, $X_{ij}^{\#} < X_{ij}^*$, where the * indicates the usage under the fixed tariff when firms don't consider the effects of losses due to an outage in their production decisions. Once these potential effects are recognized, we must redefine the CBL's (e.g., the level of usage of off-peak peak and peak electricity under the fixed tariff) as $X_{j3}^{\#}$, and $X_{i4}^{\#}$, respectively. Thus, by explicitly recognizing the possibility of an

outage in maximizing expected profit, the value of system-wide reserves to the firm is revealed, even in the case where the level of reserves is fixed. An increase in reserves will reduce the probability of an outage and/or the proportion of output lost due to the outage.[36] This will lead the firm to expand input use in anticipation of a smaller difference between expected and potential output. This result confirms the widely held belief that reliability is a critical element to the development of a robust economy, and that this is especially true of industries that require high reliability to be profitable.

5.2 Case 2: Counting Load Reductions as Additions to Reserves

Taking reliability explicitly into account, the firm's optimization problem is altered in an important way. If the firm reduces its usage below its current usage, called a CBL, when requested to do so by a reliability agent (e.g., Independent System Operator (ISO)), the load reduction is counted by the ISO as an addition to system reserves. For modeling simplicity, we assume that during off-peak times generators are able to supply sufficient electricity and reserves to meet system-wide demand and reserves and that reserve shortfalls (and potential for outages) only occur during peak periods .[37] Therefore, in our analysis, the level of reserves, R, refers to reserves on peak. In subsequent cases, we also allow for reductions in electricity use below the firm's peak CBL of $X_{4j}^{\#}$ to be counted as additional reserves.

In Case 2, the firm still takes the price of electricity and the level of system-wide reserves as supplied by the generators as given. We can count a firm's load reduction below the redefined CBL by simply adding constraints to Case 1 to illustrate situations of interest (i.e., where electricity is priced at a fixed tariff, T). We also now let the level of reserves, R, be solved for as part of the firm's constrained maximization problem. The expected profit maximization problem for the firm is now:

---

[36] The reduction in input use could be viewed as an outage loss mitigation strategy. The extent to which any firms actually consider the effect of an outage in making their input decisions is an empirical question. However, if they believe that the probability of an outage is very small and that its individual load reduction as a mitigation strategy will have a negligible effect of system reliability, the firm may not bother with making such adjustments on its own. Alternatively, this formulation explains why firms that use a lot of electricity, and forfeit a large proportion of output during an outage, are so concerned with the level of system reliability, and why they often comply, without compensation, to public appeals to reduce usage during severe reserve shortfalls to avoid a full outage.

[37] This may not characterize situation with complete accuracy due to occasional unforeseen generator outages or transmission problems.

(5.6) max. $E\Pi_j = [G^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$+ [1–G^j(R|N)]\ [1- D^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\} – P_1 X_{1j}- P_2 X_{2j} - T (X_{3j},+ X_{4j})$

subject to:

(5.7) $X_{4j} + R_j = X_{4j}^{\#}$

(5.8) $R – R_j = R^*$.

$= \{P_j F_j\}\ [\partial G^j / \partial R]$

$+ \{[1–G^j)]$ Forming the Lagrange function, we have the problem:

(5.9) max. $E'\Pi_j = [G^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$+ \mathbf{[}1–G^j(R|N\mathbf{)]}\ \mathbf{[}1- D^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\mathbf{\} – \{}P_1 X_{1j}- P_2 X_{2j} - T (X_{3j}+ X_{4j})\mathbf{\} - \{}\lambda_{x4}[X_{4j} + R_j - X_{4j}^{\#}] - \lambda_R [R – Rj - R^*]\mathbf{\}}.$

The first-order conditions are now:

(5.10) $\partial E'\Pi_j / \partial X_{1j} = \{P_j \partial F_j / \partial X_{1j}\}\{[G^j] + [1–G^j]\ [1- D^j]\} - P_1 = 0$

(5.11) $\partial E'\Pi_j / \partial X_{2j} = \{P_j \partial F_j / \partial X_{2j}\}\{[G^j] + [1–G^j]\ [1- D^j]\} - P_2 = 0$

(5.12) $\partial E'\Pi_j / \partial X_{3j} = \{P_j \partial F_j / \partial X_{3j}\}\{[G^j] + [1–G^j]\ [1- D^j]\} - T = 0$

(5.13) $\partial E'\Pi_j / \partial X_{4j} = \{P_j \partial F_j / \partial X_{4j}\}\{[G^j] + [1–G^j]\ [1- D^j]\} - T - \lambda_{x4} = 0$

(5.14) $\partial E'\Pi_j / \partial R_j = - \lambda_{x4} + \lambda_R = 0$

(5.15) $\partial E'\Pi_j / \partial R\ [-\partial D^j / \partial R] + [1- D^j]\ [-\partial G^j / \partial R]\}\{P_j F_j)\} - \lambda_R = 0.$

The first three first-order conditions are unchanged from Case one (1) above where the firm faces a fixed tariff. However, we can now solve equation (5.13), for value of the marginal product of peak electricity use, $\{P_j \partial F_j / \partial X_{4j}\}$. After rearranging equation (5.13), the value of the marginal product is now equal to the fixed tariff, plus the shadow price, in terms of expected firm profits, of peak electricity, divided by the term $\{[G^j] + [1–G^j]\ [1- D^j]\}$. That is, $\{P_j \partial F_j / \partial X_{4j}\} = [T + \lambda_{x4}] / \{[G^j] + [1–G^j]\ [1- D^j]\}$. We know that $\{[G^j] + [1–G^j]\ [1- D^j]\} < 1$, and for an interior solution with $\lambda_{x4} > 0$. For an interior solution with $\lambda_{x4} > 0$, firms will allocate some of the intended usage (CBL), $X_{4j}^{\#}$ to reserves if they are counted toward system-wide reserves by the reliability agent because doing contributes to its own welfare. .

Since the firm can now reallocate some peak electricity use to reserves, it equates the marginal value of peak electricity in production to the firm with the value of reserves

(equation (5.14)). In turn, the value of reserves is measured in terms of their contribution to expected profit (equation (5.15)). Simplifying this relationship produces:

(5.16) $\lambda_R = \{[(D^j)(\partial G^j / \partial R)] - [(1-G^j)(\partial D^j / \partial R)]\}\{P_j F_j)\} > 0$

(+) (+) (+) (-)

From this equation, it is apparent that the value of reserves is positive, and that it is determined through the difference in its two separate effects on the value of realized output. The first effect (the first term in [ ] above) is **[**(the proportion of output lost due to an outage) multiplied by (the change in the probability of no outage due to a change in reserves)**]**. The second effect (the second term in [ ]) is **[**(the probability of an outage) multiplied by the (change in the proportion of the value of output lost during an outage)**]**. Since the proportion of output lost during an outage declines as R increases, this last term is negative and the value of reserves to the firm is unambiguously positive.

Moreover, from equation (5.14), we can see that the firm equates the shadow price of an additional unit of peak CBL ($\lambda_{x4)}$, which, if it were available, could be used in production or supplied as reserves through load curtailment, to the shadow price of reserves.[38] This is a particularly convenient algebraic formulation to employed to determine the value of reserves. It makes transparent both the role of the loss of load probability and the firm's outage cost, expressed in terms of the forgone value of production.

5.3 Case: System-wide Reserves as a Public Good

In Case 3, we extend the model to account for the fact that system-wide reserves are a public good.[39] All customers enjoy the same level of system reliability due to the level of reserves, except that by indexing all functions by firm we allow for local differences in reliability facing a firm due to load pockets, etc.[40]

---

[38] This could well explain why some firms install backup generation.

[39] A public good is one in which the use of a unit of the good by one agent does not preclude its use by another. Put differently, the essential feature of a public good is that it is non-depletable (Baumol and Oates, 1988).

[40] We could also expand the model to differentiate reserves locally, but this in not normally what is done for a single electrical transmission and distribution network. However, some system are characterized by load pockets that due to transmission limits have lower reliability that the system standard. Therefore, in what follows, we ignore this possibility, recognizing that separate models involving only the load pocket could be defined to deal with that situation.

As above, we model the public good by maximizing the combined expected profits of all firms (Mas-Colell et al, 1995). The problem is now:

(5.17) max. $\sum_j E\Pi_j = \sum_j \{[G^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$+ [1–G^j(R|N)] [1- D^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$– \{P_1 \sum_j X_{1j}- P_2 \sum_j X_{2j} - T \sum_j (X_{3j}+ X_{4j})\}$

subject to:

(5.18) $X_{4j} + R_j = X_{4j}^{\#}$ (j = 1,…, M)

(5.19) $R – \sum_j R_j = R^*$.

Forming the Lagrange function, we have the problem:

(5.20) max. $E\Pi = \sum_j \{[G^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$+ [1–G^j(R|N)] [1- D^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\} – P_1 \sum_j X_{1j}- P_2 \sum_j X_{2j}$

$- T \sum_j (X_{3j}+ X_{4j}) - \sum_j \lambda_{x4j}[X_{4j} + R_j - X_{4j}^{\#}] - \lambda_R [R – \sum_j Rj - R^*]$.

The first-order conditions are now:

(5.21) $\partial E\Pi / \partial X_{1j} = \{P_j \partial F_j / \partial X_{1j}\}\{[G^j] + [1–G^j] [1- D^j]\} - P_1 = 0$, (j = 1,…, M)

(5.22) $\partial E\Pi / \partial X_{2j} = \{P_j \partial F_j / \partial X_{2j}\}\{[G^j] + [1–G^j] [1- D^j]\} - P_2 = 0$, (j = 1,…, M)

(5.23) $\partial E\Pi / \partial X_{3j} = \{P_j \partial F_j / \partial X_{3j}\}\{[G^j] + [1–G^j] [1- D^j]\} - T = 0$, (j = 1,…, M)

(5.24) $\partial E\Pi / \partial X_{4j} = \{P_j \partial F_j / \partial X_{4j}\}\{[G^j] + [1–G^j] [1- D^j]\} - T - \lambda_{x4j} = 0$, (j = 1,…, M)

(5.25) $\partial E\Pi / \partial R_j = - \lambda_{x4j} + \lambda_R = 0$, (j = 1,…, M)

(5.26) $\partial E\Pi / \partial R = \{P_j F_j\} [\partial G^j / \partial R]$

$+ \{[1–G^j)] [-\partial D^j / \partial R] + [1- D^j] [-\partial G^j / \partial R]\}\{P_j F_j)\} + \sum_{i \neq j} \{\{P_i F_i\} [\partial G^i / \partial R]$

$+ \{[1–G^i)] [-\partial D^i / \partial R] + [1- D^i] [-\partial G^i / \partial R]\}\{P_i F_i)\}\} - \lambda_R = 0$.

What is immediately evident about this set of first-order conditions is that there is a separate condition for each of the four inputs for each of the M firms (5.21-5.25), but they are identical to conditions in Case 2 for an individual firm maximizing expected profits. With the exception of peak electricity, each firm still equates the value of the marginal product of an input with its price divided by $\{[G^j] + [1–G^j] [1- D^j]\}$. For peak electricity, the value of the marginal product is equated to its price plus its shadow value (equation (5.24)) due to the fact that peak electricity use plus load relief must be equal to the firm's peak CBL.

This shadow value is set equal to the value of reserves (equation (5.25). When viewed from a public-good perspective, the true, system-wide value of reserves is apparent from equation (5.26). Viewed from the perspective of any one firm j, the value of reserves includes the value of reserves' two separate effects on firm j's expected output. This is the first term in equation (5.27) below, which is equal to the right-hand side of equation (5.16) above. There are now M-1 similar terms as shown below in equation (5.27) reflecting the sum of the value of system-wide reserves' two separate effects on the value of its own and other firms' realized output:

$$(5.27)\ \lambda_R = \{[(D^j)(\partial G^j / \partial R)] - [(1-G^j)(\partial D^j / \partial R)]\}\ \{P_j F_j)\}$$
$$(+)\quad (+)\qquad (+)\quad (-)$$
$$+ \sum_{i \neq j} \{[(D^i)(\partial G^i / \partial R)] - [(1-G^i)(\partial D^i / \partial R)]\}\ \{P_i F_i)\} > 0$$
$$(+)\quad (+)\qquad (+)\quad (-)$$

As before, for each firm, the first effect (the first term in [ ] in (5.27) above) is:

> **[**(the proportion of output lost due to an outage) multiplied by (the change in the probability of no outage due to a change in reserves**)]**,

and this term is unequivocally positive. The second effect (the second term in [ ]) is:

> **[**(the probability of an outage) multiplied by the (change in the proportion of the value of output lost during an outage)**]**.

This expression is also positive and larger than the equivalent firm in equation (5.16), the socially optimal level of the firm's peak electricity use is smaller than if each firm individually maximizes expected profits. Consequently, the level of reserves would now be higher.[41]

[41] It is tempting to conclude that by comparing the results of case 2 with case 3, we can establish the inefficiency or market failure due to the public nature of reserves. However, such a comparison is not strictly valid from one important perspective. In case 2, we assumed that a firm accounts only for its own benefits of supplying reserves in its optimization problem. However, we have not accounted for the fact that all other firms do likewise. Therefore, it is likely that each firm will supply some small amount of reserves, and to show that the total supply is still too small from society's point of view, we must compare case 3 with a situation in which each firm makes its profit maximizing decisions (including its level of reserves supplied) based on the combined levels of reserves supplied individually by other firms. Mas-Colell et al. (1995) develops a similar proof for a consumer problem. In maximizing profits (utility) the firm (consumer) would take as given the amount of the pubic good used (purchased) by all other firms (consumers). One could think of these decisions on how much of the public good to use (purchase) as being determined in the same way as they would be in determining a Nash equilibrium. All that is then required to show that private actions are not socially optimal is for at least two firms (consumers) to benefit in some

The critical difference between this result and the one for Case 2 comes from equation (5.25). Viewed from this social perspective, each firm should now equate the shadow price of an additional unit of its peak CBL (which could be used in production or supplied as reserves through load curtailment) to this now much larger shadow price of reserves. Put differently, at the optimal level of public-good reserves, we know that the sum of the firms' marginal benefits (in terms of reduced expected outage costs) should be set equal to the marginal cost of supplying reserves. This is in stark contrast to the situation in Case 2 where each individual firm supplies reserves only to the point where its own marginal expected benefit is equal to expected cost.

Left to their own devices, however, firms will not voluntarily curtail load to the point that the marginal value product of peak energy use reaches the combined value of reserves to all firms. The failure of each firm to consider the benefits to other firms in its decision to supply reserves is the standard free-rider problem. Each firm has an incentive to enjoy the benefits of the public good supplied by others, while it supplies an insufficient amount.[42] But, if all firms think that way, none supplies more than the level that equates its internalized value with the internalized reliability benefit, and the socially optimal amount of curtailment is not forthcoming met. As a result, the greater social good is never achieved without perceptive intervention; somebody that to recognize that there are gains form trading among firms.

From society's perspective, the industry demand curve for reserves can now be thought of geometrically as the *vertical* summation of individual firm's demand curves, in contrast to the *horizontal* sum of demand curves to obtain the industry demand for a private good. This problem of market failure caused by the undersupply of reserves through reliance of the private provision of the public good can be resolved through market intervention. For example direct quantity intervention (such as governmental provision) or through "market-based" taxes or subsidies perhaps through government financed investment by firms in self-generation capacity in amounts justified by their private valuation shortfall. Some advocate imposing ICAP requirements that extend

---

way from the use (consumption) of public good. This is sufficient for condition (5.25) to diverge from (5.36) even when other firms' private supply of reserves is included.

[42] In a stylized model of this kind it is easy to show that it is only the firm with the largest marginal benefits (reduction in expected outage costs) that will supply any reserves through load curtailment. This result can be derived in a similar fashion to the consumer example in Mas-Colell (1995, p. 362).

several years into the future as a means of bringing forth the desired amount of reliability, which seems like a step back to the regulated monopoly days where markets were administered.

An alternative, and compelling way to accomplishing the desired curtailment behavior by firms is to pay them up to the full social value. Accordingly, one would offer an incentive to some firms for the provision of system-wide reserves up to the level equal to the marginal benefits to all firms the additional supply reserves. Any greater incentive would reduce social benefits, and result in lower benefits that are achievable. The amount of this incentive is given in equation (5.27). However, if firms can be enticed to curtail for less than the full social value, then there will be a net social gain.

There are clearly a number of practical issues that need to be addressed in implementing such a subsidy scheme for internalizing the public-good benefits of system-wide reserves. Perhaps the most important issue relates to the fact that in order to establish the appropriate level of incentive to offer for curtailments, the public agency must know the benefits derived from reserves by each of the firms. This is essentially their willingness to pay for reserves, as measured through each firm's outage cost. Unlike the case for a private good, there clearly is no market mechanism by which this willingness to pay is revealed. However, the maximum that a firm would be willing to pay to avoid an outage is clearly given in this model by the fraction of the value of production lost during an outage. In addition to firm-level estimates of outage costs, the calculation of subsidy rates must be based on an estimated loss of load probability function (LOLP), which is given by $[1 - G(R|N)]$. None of these is easily quantified.

We next explore two cases in which conditions change slightly if the actions of firms can affect electricity price and the price of reserves. The changes are primarily in terms of how the costs of supplying reserves from other sources are affected.

5.4 Case 4: Load Reduction Can Affect the Price of Electricity

As an extension of Case 3, we now recognize that the actions of these firms can affect the price of electricity by their decisions to use less electricity than the CBL under condition specified by the ISO. The availability of reserves from generators and other

traditional sources is still fixed at R^*; the only additions to reserves in this model come from firms' load curtailments.

We need to substitute an inverse supply curve for both off-peak and peak electricity ($P_{X4} = S_4(X_4)$) and ($P_{X3} = S_3(X_3)$) for the price of electricity, [43] and we assume that $\partial S/\partial X_i > 0$, for i = 3 and 4; both supply curves are upward sloping. Finally, we add constraints to ensure that the aggregate demand for peak and off-peak electricity is equal to the sum of the firms' individual demands, $X_4 = \sum_j X_{4j}$ and $X_3 = \sum_j X_{3j}$. The profit maximization challenge firm face is:

(5.28) max. $E\Pi = \sum_j \{[G^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$+ [1-G^j(R|N)] [1- D^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\} - P_1 \sum_j X_{1j} - P_2 \sum_j X_{2j}$

$- S_3(X_3) X_3 - \lambda_3 [\sum_j X_{3j} - X_3] - S_4(X_4) X_4 - \lambda_4 [\sum_j X_{4j} - X_4] - \sum_j \lambda_{x4j}[X_{4j} + R_j - X_{4j}^{\#}]$

$- \lambda_R [R - \sum_j Rj - R^*]$.

The first-order conditions are now:

(5.29) $\partial E\Pi / \partial X_{1j} = \{P_j \partial F_j / \partial X_{1j}\}\{[G^j] + [1-G^j] [1- D^j]\} - P_1 = 0$, (j = 1,…,M)

(5.30) $\partial E\Pi / \partial X_{2j} = \{P_j \partial F_j / \partial X_{2j}\}\{[G^j] + [1-G^j] [1- D^j]\} - P_2 = 0$, (j = 1,…,M)

(5.31) $\partial E\Pi / \partial X_{3j} = \{P_j \partial F_j / \partial X_{3j}\}\{[G^j] + [1-G^j] [1- D^j]\} - \lambda_3 = 0$, (j = 1,…,M)

(5.32) $\partial E\Pi / \partial X_{4j} = \{P_j \partial F_j / \partial X_{4j}\}\{[G^j] + [1-G^j] [1- D^j]\} - \lambda_4 - \lambda_{x4j} = 0$, (j = 1,…,M)

(5.33) $\partial E\Pi / \partial X_3 = - [(\partial S_3 / \partial X_3)(X_3) + S_3(X_3) (\partial X_3 / \partial X_3)] + \lambda_3 = 0$

(5.34) $\partial E\Pi / \partial X_4 = - [(\partial S_4 / \partial X_4)(X_4) + S_4(X_4) (\partial X_4 / \partial X_4)] + \lambda_4 = 0$

(5.35) $\partial E\Pi / \partial R_j = - \lambda_{x4j} + \lambda_R = 0$, (j = 1,…,M)

(5.36) $\partial E\Pi / \partial R = \{P_j F_j\} [\partial G^j / \partial R]$

$+ \{[1-G^j)] [-\partial D^j / \partial R] + [1- D^j] [-\partial G^j / \partial R]\}\{P_j F_j)\} + \sum_{i \neq j} \{\{P_i F_i\} [\partial G^i / \partial R]$

$+ \{[1-G^i)] [-\partial D^i / \partial R] + [1- D^i] [-\partial G^i / \partial R]\}\{P_i F_i)\}\} - \lambda_R = 0$.

Because firms' decisions to purchase electricity affect prices, we now have two additional first-order conditions: the "outage probability weighted" marginal value product of electricity in each period to each firm, which is at least equal to initial prices,

[43] As noted above, it remains convenient to treat the electricity supply side of this analysis as consisting of a single profit maximizing generation firm whose off-peak and peak cost functions are $C_3(X_3)$ and $C_4(X_4)$, and marginal cost functions given by $S_3(X_3)$ and $S_4(X_4)$, respectively. Mas-Colell (1995, p.147-48) demonstrates that this is equivalent to the supply behavior of an industry composed of J price-taking firms whose aggregate cost function are given by $C_3(X_3)$ and $C_4(X_4)$, respectively.

plus the effect at the margin of changes in aggregate demand on price (equations (5.33) and (5.34)). We can also be assured that the market clearing peak (off-peak) price is above (below) that of the case of a fixed tariff, T. We would expect electricity demand in off-peak periods to rise, partially (or totally) offsetting the reduction in use due to the explicit consideration of the effect of an outage on expected profit. We know that the new demand is above the off-peak CBL of $X_{3j}^{\#}$ (defined in Case 1 of Section 5).

Whether usage is above or below the CBL when there is no explicit recognition of an outage by the firm, X_{3j}* (from Section 2) is an empirical question, and it depends on the relative strength of the two opposing effects. Since the market-clearing price for peak electricity will be above T, this effect on peak price will only serve to reinforce the incentives to reduce demand due to recognizing the probability of an outage and the public good value of reserves. Thus, peak demand will be unequivocally below both X_{4j}* (the level under the fixed tariff) and $X_{4j}^{\#}$ (the level when the firm only considers the value of reliability to itself).

5.5 Case 5: Generators Can Also Supply Additional Reserves

In Case 5, we analyze the situation in which generators are allowed to supply reserves over and above the fixed level assumed in the above cases, in effect pitting then against the curtailment bids of firms. The actions of both customers and generators now affect the price of reserves and the price of electricity. For this case, we must specify an inverse supply curve for reserves from the generators, call it $A(R^E)$, where $\partial A/\partial R^E > 0$, and require that total reserves are now equal to what is supplied by generators as well as that supplied through load relief.[44] Finally, we must ensure that electricity use plus reserves from all sources during the peak period are no greater than the total capacity of the generators C^{RE}. The problem is:

(5.37) max. $E\Pi = \sum_j \{[G^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\}$

$+ [1-G^j(R|N)]\,[1- D^j(R|N)]\{P_j F^j(X_{1j}, X_{2j}, X_{3j}, X_{4j})\} - P_1 \sum_j X_{1j} - P_2 \sum_j X_{2j}$

$- S_3(X_3)\, X_3 - \lambda_3 [\sum_j X_{3j} - X_3] - S_4(X_4)\, X_4 - \lambda_4 [\sum_j X_{4j} - X_4] - \sum_j \lambda_{x4j}[X_{4j} + R_j - X_{4j}^{\#}]$

[44] We can think of this supply curve for reserves from generators being derived in the same way as the supply curves for the peak and off-peak electricity in case 4 above.

$$- A(R^E)\,(R^E) - \lambda_R\,[R - \textstyle\sum_j R_j - R^E] - \lambda_G\,[X_4 + \textstyle\sum_j R_j + R^E - C^{RE}]$$

The first-order conditions are now:

(5.38) $\partial E\Pi / \partial X_{1j} = \{P_j\, \partial F_j / \partial X_{1j}\}\{[G^j] + [1-G^j]\,[1- D^j]\} - P_1 = 0,\ (j = 1,\ldots, M)$

(5.39) $\partial E\Pi / \partial X_{2j} = \{P_j\, \partial F_j / \partial X_{2j}\}\{[G^j] + [1-G^j]\,[1- D^j]\} - P_2 = 0,\ (j = 1,\ldots, M)$

(5.40) $\partial E\Pi / \partial X_{3j} = \{P_j\, \partial F_j / \partial X_{3j}\}\{[G^j] + [1-G^j]\,[1- D^j]\} - \lambda_3 = 0,\ (j = 1,\ldots,M)$

(5.41) $\partial E\Pi / \partial X_{4j} = \{P_j\, \partial F_j / \partial X_{4j}\}\{[G^j] + [1-G^j]\,[1- D^j]\} - \lambda_4 - \lambda_{x4j} = 0,\ (j = 1,\ldots,M)$

(5.42) $\partial E\Pi / \partial X_3 = - [(\partial S_3 / \partial X_3)(X_3) + S_3(X_3)\,(\partial X_3 / \partial X_3)] + \lambda_3 = 0$

(5.43) $\partial E\Pi / \partial X_4 = - [(\partial S_4 / \partial X_4)(X_4) + S_4(X_4)\,(\partial X_4 / \partial X_4)] + \lambda_4 - \lambda_G = 0$

(5.44) $\partial E\Pi / \partial R^E = - [(\partial A / \partial R^E)(R^E) + A(R^E)(\partial R^E / \partial R^E)] + \lambda_R - \lambda_G = 0$

(5.45) $\partial E\Pi / \partial R_j = - \lambda_{x4j} + \lambda_R - \lambda_G = 0,\ (j = 1,\ldots,M)$

(5.46) $\partial E\Pi / \partial R = \{P_j\, F_j\}\,[\partial G^j / \partial R]$

$+ \{[1-G^j)]\,[-\partial D^j / \partial R] + [1- D^j]\,[-\partial G^j / \partial R]\}\{P_j\, F_j)\} + \textstyle\sum_{i \neq j} \{\{P_i\, F_i\}\,[\partial G^i / \partial R]$

$+ \{[1-G^i)]\,[-\partial D^i / \partial R] + [1- D^i]\,[-\partial G^i / \partial R]\}\{P_i\, F_i)\}\} - \lambda_R = 0.$

As one might expect, the first-order conditions for this model are quite similar to the ones for Case 4. In fact, the first five conditions are identical. Further, from equation (5.46) we see that the shadow price of reserves is set in the same way as in Case 4. However, equation (5.43) differs from equation (5.34), its counterpart in Case 4 above. There is a similar difference between equations (5.44) and (5.35). The new peak electricity price is no longer set equal to the shadow price of peak electricity, but it is set equal to the difference between the shadow price of peak electricity less the shadow price of generation capacity ($\lambda_4 - \lambda_G$). The important result is that the price an additional unit of CBL, which is equivalent to another unit of load reduction reserves available, to any firm is now set equal to the difference between the shadow price of reserves and the shadow price of generation ($\lambda_R - \lambda_G$). Since the constraint on capacity ensures that electricity use plus reserves from all sources during the peak period are no greater than the total capacity of the generators C^{RE}, the complementary slackness conditions guarantee that $\lambda_G = 0$ unless the constraint binds. Thus, if there is excess capacity, the price of peak electricity and the price of load reduction reserves are equal to the respective shadow prices, as they were in Case 4 above.

We can gain further insights into the value of peak electricity and reserves by solving both (5.43) and (5.44) for λ_G and equating the two expressions. Thus, we have:

(5.47) $\lambda_4 - [(\partial S_4 / \partial X_4)(X_4) + S_4(X_4)(\partial X_4 / \partial X_4)] = \lambda_R - [(\partial A / \partial R^E)(R^E) + A(R^E)(\partial R^E / \partial R^E)]$.

Rearranging, we have:

(5.48) $\lambda_R - \lambda_4 = [(\partial A / \partial R^E)(R^E) + A(R^E)(\partial R^E / \partial R^E)] - [(\partial S_4 / \partial X_4)(X_4) + S_4(X_4)(\partial X_4 / \partial X_4)]$

When sufficient reserves are available, the difference in the value of reserves from load reduction and from generation is equated to the difference in the additional cost of obtaining reserves from these additional sources.[45] Thus, under normal system conditions, when supply and reserves are sufficient, the difference in value would be captured in the equilibrium price differences. Under these conditions, the probability of an outage would be very low, and the value of additional reserve, as measured by the system wide expected outage costs from equation (5.56), would be very low. Reserves in the form of load reduction would command no premium in the market relative to reserves supplied from generators, even if the public-good nature of reserves is recognized explicitly. This is a key result, as it says there is no subsidy, implicit or explicit, in the valuation of load reductions relative to what generators would be paid.

However, suppose that during an emergency situation there is a sudden reserve shortfall, perhaps through an unexpected generator outage or an episodic rise in demand. Further, suppose that there are no additional generator-supplied peak energy or reserves to dispatch, essentially creating a disequilibrium situation.[46] Under these conditions, the probability of an outage $[1-G^i]$ (for all i firms) would rise, in the worst cases dramatically. The value of lost load (VOLL) could rise accordingly, as reflected by the increase in the proportion of a firm's output lost due to the outage (D^i). Thus, this disequilibrium situation could lead to a substantial increase in the combined expected outage costs of all firms given in equation (5.46). The social value of reserves would rise accordingly, which

[45] If $(\partial S_4 / \partial X_4)$ and $(\partial A / \partial R^E)$ from equation (5.48) were written in percentage terms, they would represent the supply flexibilities for the supply of peak electricity and generator supplied reserves, respectively.

[46] This disequilibrium situation might well be reflected by the supply curves for generator supplied reserves and energy becoming vertical, or nearly so. This means that both $(\partial S_4 / \partial X_4)$ and $(\partial A / \partial R^E)$ would become very large.

now represents the amount that one could afford to pay firms for the load reduction (e.g. additional reserves) needed to restore system reliability back to its design level (Stoft, 2002). But, since there is no generation available, the situation can only be resolved by dispatching curtailments.

The essential conclusion is that under these disequilibrium situations, the value of load reduction is determined by combined expected outage costs, and it bears no explicit relationship to the market-clearing prices for energy or generator-supplied reserves either prior to the emergency or after system reliability has been restored.[47] It could be the case that if the firms' combined VOLLs were high, the "budget" for curtailments would not be sufficient to restore reserves (Stoft, 2002).

6. Policy Implications

Despite the growing enthusiasm for the new generation of demand response (DR) programs, which include real time pricing (RTP) and ISO/utility Demand Response (DR) programs, there has been limited and highly abstract discussion in the literature to demonstrate exactly how these programs can contribute to market efficiency, the management of market risk, and overall social welfare. By examining both the market for energy and reserves separately and in combination, this paper sets out a foundation for evaluating incentives and the expenditure of public funds to promote retail customer price responsiveness to achieve the welfare benefits that result.

We recognize that if demand response were to come about spontaneously, or by regulatory fiat, with the onset of retail competition, then the social good would be served naturally, but necessarily fully. However, we are skeptical that the full and necessary amount of DR will immediately become apparent and active on its own. Consequently we turn attention to the question of what, if any, incentives would be effective and are warranted to induce demand response by retail customers.

New competitive electricity markets seek to achieve these welfare gains by exposing load serving entities (LSE) and other commodity providers to hourly prices, which, in theory at least, should approach marginal costs and could differ dramatically

[47] This suggests that these resources should not set the marginal market price as these situations go beyond what can be construed as scarcity pricing.

between peak and off-peak periods. However, we show that unless retail customers also see these prices, as opposed to conventional fixed tariffs, and adjust their demands accordingly, unnecessary and potentially large deadweight losses will persist. Our analysis focused initially only on the market for electric energy and described a competitive electricity market where demand can fully adjust to price. We then described the welfare losses that occur in current wholesale electricity markets because the preponderance of retail customers face fixed tariffs, eliminating any incentive to respond to contemporaneous price volatility. Finally, we showed that some of these welfare losses that arise because of the lack of price responsiveness can be eliminated by introducing ISO-initiated demand response (DR) programs in which at least some customers are paid to reduce load when prices are high. If done properly, the result is the achievement of desirable welfare gains.

Some DR programs require that customers submit bids for load reduction; load curtailments are scheduled when the market price exceeds the customer's bid. Examples are the day-ahead market DR programs of NYISO and PJM. An alternative is a program whereby customers face market prices for commodity and adjust their usage in response to price based on the bill saving they would realize—they receive an implicit rather than an explicit curtailment payment. The Niagara Mohawk SC-3A default service rate for large customers is the first and most prominent example of this service. In the latter case, the customer is exposed to the full force of market price volatility, while the former allows the customer to choose under what circumstances it faces market prices.

A third design allows customers to hedge their typical usage (called a CBL) under a standard tariff but settle hourly deviations from the CBL at prices that reflect prevailing supply costs that are posted a day in advance. The first of these programs was implemented by Niagara Mohawk in the late 1980s and followed by a few highly prominent applications at Georgia Power, Duke Power, and Public Service of Oklahoma among others. Critical peak pricing programs operated by vertically integrated utilities are a variation on this theme whereby the utility is allowed to override the usual rate with one much higher under certain, well specified conditions.

If the three program designs achieve similar levels of participation and curtailments under similar market circumstances, then the one that involves the least

incentives would be preferable. That would be the full exposure model implemented by Niagara Mohawk since it involves no incentive payment or concession for the reduction in deadweight losses.[48] The other designs involve some form of incentive payment, which must be weighted against the reduced deadweight losses other customers realize.

And, all of these programs result in transfers of revenue from suppliers to either customers or the customer's load serving entity as long as the curtailments that are undertaken result in a reduction in market prices from the level it otherwise would have achieved. Our graphical analyses in Section 2 illustrate how curtailments induced by high prices, especially when supply is tight and the supply curve very inelastic, result in a drop in the market-clearing price that results in lower expenditures for both the curtailing customers and all other customers facing those prices or their consequences. Economists treat such transfers as inconsequential since they have no way of normalizing the impacts on winners and losers in a way that allows for a meaningful comparison. However, policy makers are not so indifferent, and neither are many market participants.

The challenge facing policymakers is how to design retail programs that can eliminate or reduce the size of the deadweight losses in wholesale electricity markets, and address the equity issues implied by the transfers that inevitable result. Given the diversity of customers, particularly their varying abilities to adjust load in response to price, there is perhaps no single solution to the problem. In this paper, we have focused on illustrating the potential welfare effects of a Demand Response (DR) program where firms bid to be paid to reduce load if prices exceed their bid prices. We demonstrate graphically that the conditions under which payments for load reduction under "economic" DR programs would be less than the welfare gains clearly depends on the relative slopes (or elasticities) of the supply and demand curves for electricity.

Viewed from this perspective, it is clear that as the supply curve becomes steeper, *ceteris paribus*, the net welfare gains would seem to justify paying inducements to customers to curtail DR load. Similarly, the less price-responsive (steeper) the initial demand curve, *ceteris paribus*, the larger the net welfare gains from load reductions. In summary, the potential welfare gains from DR programs are highest in markets where

[48] This is the closest thing to a natural demand response program since customers face the market price for all commodities as metered.

both the supply and demand curves are initially extremely price inelastic (e.g., the "steeper" both curves are), which is the case during at least a few hours of the year.

This view provides a basis for understanding the size of the deadweight losses and the potential gains from implementing DR programs. In the case of electricity markets, the realization of welfare gains is further complicated because many customers are likely to be unwilling or unable to continually adjust electricity usage, because the transactions costs of doing so outweigh the potential benefits. A relatively small number of customers, on the other hand, may be able to profitably adjust demand to real time prices. For a third group, those customers that respond to price only on a limited basis, efficiency gains may be realized through rates that compensate customers for load reductions when prices are extremely high and/or when system reliability is threatened, a hybrid of the two extreme cases.

It is difficult to classify customers into these groups because of the paucity of data on observed price responses. Moreover, customers may need some incentive to undertake analyses or experiment with a DR service in order to self-classify themselves by their ability to be price responsive. Thus, public policy involvement can be justified solely to help identify or estimate the potential price responsiveness of customers, and to educate customers about how to respond to price, thus providing a knowledge base that marketers can use to match service offerings to meet the needs of a diverse set of customers. If such programs are successful, customers may ultimately adjust usage such that they no longer require incentives to curtail; the incentive would be built into the negotiated service contract.

We expand the analysis of the electricity market by accounting explicitly for the supply and demand for both the electricity commodity and reserves. Reserves are additional resources, above what is needed to balance energy usage, that are committed for the purposes of maintaining system reliability at acceptable levels.

There are two unique features of the model used to conduct this expanded analysis. First, the customer's demand for reserves, by reducing the probability of losses due to an outage, is modeled as a damage control agent rather than a conventional input. This damage control agent reduces both the probability of power outages and the firms' economic losses by reducing the difference between planned output and output that

would actually be realized in the event of an outage. Further, the function that relates reserves and system states to the probability of an outage (loss of load probability--LOLP) is specified separately from the function that relates reserves and system states to a firm's losses during an outage. For this model, firms are now assumed to maximize expected profits.

Second, we also recognize explicitly that reserves are a public good in the sense that all customers are provided the same level of reserves due to the common transmission and distribution system. This public good nature of reserves is clearly another rationale for considering what policy actions are warranted to ensure that customer's valuation of service is incorporated into new competitive electricity markets. By modeling five increasingly complex situations within this framework, we are able to isolate the effects of different actions by customers and generators.

Case 1. The firm maximizes expected profit for a given level of system reserves.

Case 2. The firm is allowed to reduce load below its CBL; this load reduction is recognized by the ISO as an addition to system reserves.

Case 3. The public good nature of the system-wide reserves is considered explicitly.

Case 4. The combined load curtailments of firms are assumed to affect electricity price.

Case 5. Generators can supply reserves over and above the fixed level in Case 1; customers can affect prices of reserves and electricity.

Using Case 1, we verify the graphic results from Section 2. But, once load reductions can be counted as reserves, firms reduce usage because they find it to their advantage to increase system reliability even if they can only capture the "private" benefits to themselves from this enhanced system reliability. As one would expect, these independent actions by firms fall well short of the optimal level of load reduction from society's point of view. However, once the public-good nature of reserves is recognized explicitly, the value of any single firm's load reduction potentially increases dramatically; it is now worth the *combined* value of the *combined* reduction in expected outage costs due to the added reserves. But, how does an individual customer collect the rents that others realize when it acts unilaterally?

When there are sufficient reserves available, the difference in the value of reserves from load reduction and that available from generators is equated to the

difference in the additional cost of obtaining reserves from these additional sources. Under normal system conditions, one could well expect this difference to be very small—the effect of a unit increase in the supply of reserves through load reduction or generator-supplied reserves would have an extremely modest effect on the price of peak energy or the price of generator-supplied reserves. Under such *normal* conditions, the probability of an outage would be very low indeed, and the value to the system of additional reserves (as measured by the system wide expected outage costs for all firms), would be extremely small as well. Reserves in the form of load reduction would command no unnecessarily high premium in the market relative to reserves supplied from generators, even when the public-good nature of reserves is recognized explicitly.

However, under an emergency situation, where there is a sudden reserve shortfall (perhaps through an unexpected generator outage, an unexpected rise in demand, etc.) and there are no more additional generator-supplied peak energy or reserves to dispatch, the situation is quite different. Under this disequilibrium situation, LOLP for all firms would rise dramatically, as would the proportion of firms' output lost due to the outage. This could in turn give rise to substantial increase in the combined expected outage costs of all firms. The social value of the load reduction (e.g. in the form of additional reserves) needed to restore system reliability would rise accordingly. The value of load reduction is determined by combined expected outage costs of all firms (the combined expected value of un-served energy—EVUE). This value is the maximum the system can afford to pay for these "load reduction" reserves and it has particular policy significance since EVUE under these conditions bears no necessary relationship to the prices for energy or generator-supplied reserves, either prior to the emergency or after system reliability has been restored.

Consequently, market prices cannot be used solely as a guide to setting the price for load curtailments needed during emergency situations to restore system reliability to design levels. Here, the issue of market rights also seems to be much clearer than in comparing DR and RTP programs in the markets for electric energy. In this case, payments for load reduction are being made to purchase a public good needed to restore system security. There is no alternative supply of the good, and payments are based on

the expected value of the combined losses avoided, so the imperative is to find ways to mitigate the impacts of outages.

It is useful to put these results into somewhat of a historical perspective. Prior to the existence of competitive electricity markets, a vertically integrated utility could use conventional DR programs to leverage its *linkage* between capacity investment decisions, based on forecasts of native load needs, and its dispatch of available units to supply customers' exigent load requirements. As we have demonstrated, these programs could be justified in terms of the substantial benefits to all stakeholders by allowing utilities to defer investments in additional capacity needed to meet established reliability criteria. When an RTO assumes responsibility for reliability for customers connected to its enfranchised grid, this important *linkage* is severed.

While individual utilities retain obligations to serve their customers, the RTO dispatches all units connected to the grid in order to meet the greater, aggregate supply and reliability obligations. The justification for utilities to contract with customers for curtailable loads is undermined, and the programs are endangered. But, as RTOs assume responsibility for electricity networks where these "peaking resources" have been an integral part of the supply planning of member utilities, the loss of these important "demand-side" resources threatens the RTO's ability to maintain reliability of the larger system. Moreover, customers lose their only point of access to wholesale markets.

Why should RTOs be concerned with load management programs? None of the new electricity markets has been immune to substantial price variability. There is mounting evidence that electricity supply curves in competitive markets rise abruptly as demand nears the system's peak; serving small increments of additional load can lead to extreme price spikes. These are exactly the conditions under which the net social welfare gains to DR programs could be the largest. On the flip side, small reductions in load, of the magnitude that might well be forthcoming through price-induced load curtailment programs, could very well abate these price spikes and reduce overall price variability. Put differently, when properly aligned with market conditions, a small amount of load management could go a long way in protecting system reliability under emergency conditions, and would also serve to discipline market prices on an ongoing basis. It is

clearly in the public interest to facilitate the design and implementation of these programs that extend market access to retail customers.

Acknowledgements

The authors are indebted to Charles Goldman for the insight and inspiration that he contributed to research that is reported herein. The authors would like to thank Steven Braithwaite and Joseph Eto for their helpful review comments. The authors are solely responsible for any errors or omissions contained in this report.

Work reported here was coordinated by the Consortium for Electric Reliability Technology Solutions (CERTS) and funded by the Assistant Secretary of Energy Efficiency and Renewable Energy, Distributed Energy and Electricity Reliability Program, Transmission Reliability, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

References

Arrow, K., H. Chenery, B. Minhas, and R. Solow 1961. "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics 43,* August, pp.225-50.

Baumol, W. and W. Oates 1988. *The Theory of Environmental Policy*. (2nd ed.) Cambridge: Cambridge University Press.

Beattie, B. and C. Taylor 1985, *The Economics of Production,* New York: John Wiley & Sons.

Berndt, E. 1991. *The Practice of Econometrics: Classic and Contemporary*. Reading MA: Addison-Wesley Publishing Company.

Boisvert, R., T. Schmit, and A. Regmi 1997. "Spatial, Productivity, and Environmental Determinants of Farmland Values," *American Journal of Agricultural Economics*. 79, pp. 1657-64.

Boisvert, R. P. Cappers, and B. Neenan 2002. "The Benefits of Customer Participation in Wholesale Electricity Markets". *The Electricity Journal,* April.

Borenstein, S. and S.P. Holland 2002. "Investment Efficiency in Competitive Electricity Markets with and without Time-Varying Retail Prices," CSEM WP 106, University of California Energy Institute, UC Berkeley, Berkeley, CA.

Caves *et al.* 2000."Mitigating Price Spikes in Wholesale Markets through Market-Based Pricing in Retail Markets." *The Electricity Journal,* April.

Ferguson, C.E. 1969. *The Neoclassical Theory of Production and Distribution,"* Cambridge: The Cambridge University Press.

Fox, G. and A. Weersink 1995. "Damage Control and Increasing Returns". *American Journal of Agricultural Economics*, 77(February) pp. 33-39.

Henderson, J. R. Quandt 1980. Microeconomic Theory: A Mathematical Approach. 3rd ed. New York McGraw-Hill Book Company.

Just, R., D. Hueth, and A. Schmitz 1982. *Applied Welfare Economics and Public Policy*, Englewood Cliffs, N. J.: Prentice Hall, Inc.

Miller, M., N. Whittlesey, and T. Barr 1975. "The Constant Elasticity of Substitution Production Function and Its Application in Research". Technical Bulletin 80, College of Agricultural Research Center, Washington State University, Pullman WA.

Litchenburg and D. Zilberman 1986. "The Econometrics of Damage Control: Why Specification Matters." *American Journal of Agricultural Economics*, 68(May) pp. 261-73.

Neenan, B. R. Boisvert, and P. Cappers 2002. "What Makes a Customer Price Responsive?" *The Electricity Journal*, April.

Mas-Colell, A., M. Whinston, and J. Green 1995. *Microeconomic Theory*, New York: Oxford University Press.

Moroney, R. J. 1972. *The Structure of Production in American Manufacturing*, Chapel Hill: The University of North Carolina Press.

Ruff, L. E. 2002. "Economic Principles of Demand Response in Electricity," Edison Electric Institute. September 3.

Schwarz, P., T. Taylor, M. Birmingham, and S. Dardan 2002. "Industrial Response to Real-Time Prices for Electricity and Utilities". *Economic Inquiry*, forthcoming.

Stoft, S. 2002. *Power System Economics: Designing Markets for Electricity*, IEEE Press, Wiley-Interscience, New York: John Wiley & Sons, Inc.

Spulber, D. 1985. "Effluent Regulation and Long-Run Optimality". *Journal of Environmental Economics and Management* 12, pp.103-116.

Uzawa, H. 1962. "Production Functions with Constant Elasticities of Substitution.." *Review of Economic Studies*, 39, pp.291-9.

Figure 1. Net Welfare Gain from PRL Programs in Competitive Electricity Markets

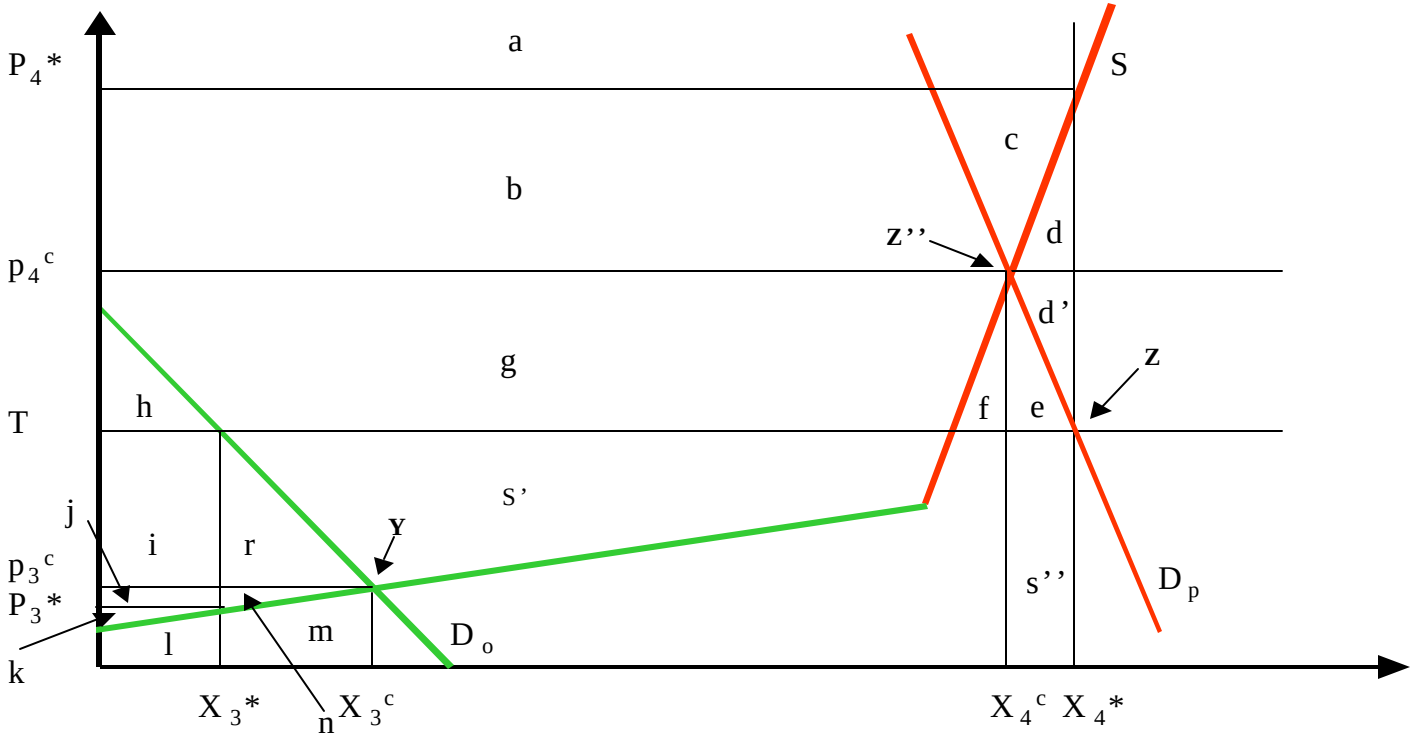


Figure 2. Net Welfare Gain, Price-Cap Load PRL Program in Competitive Electricity Markets

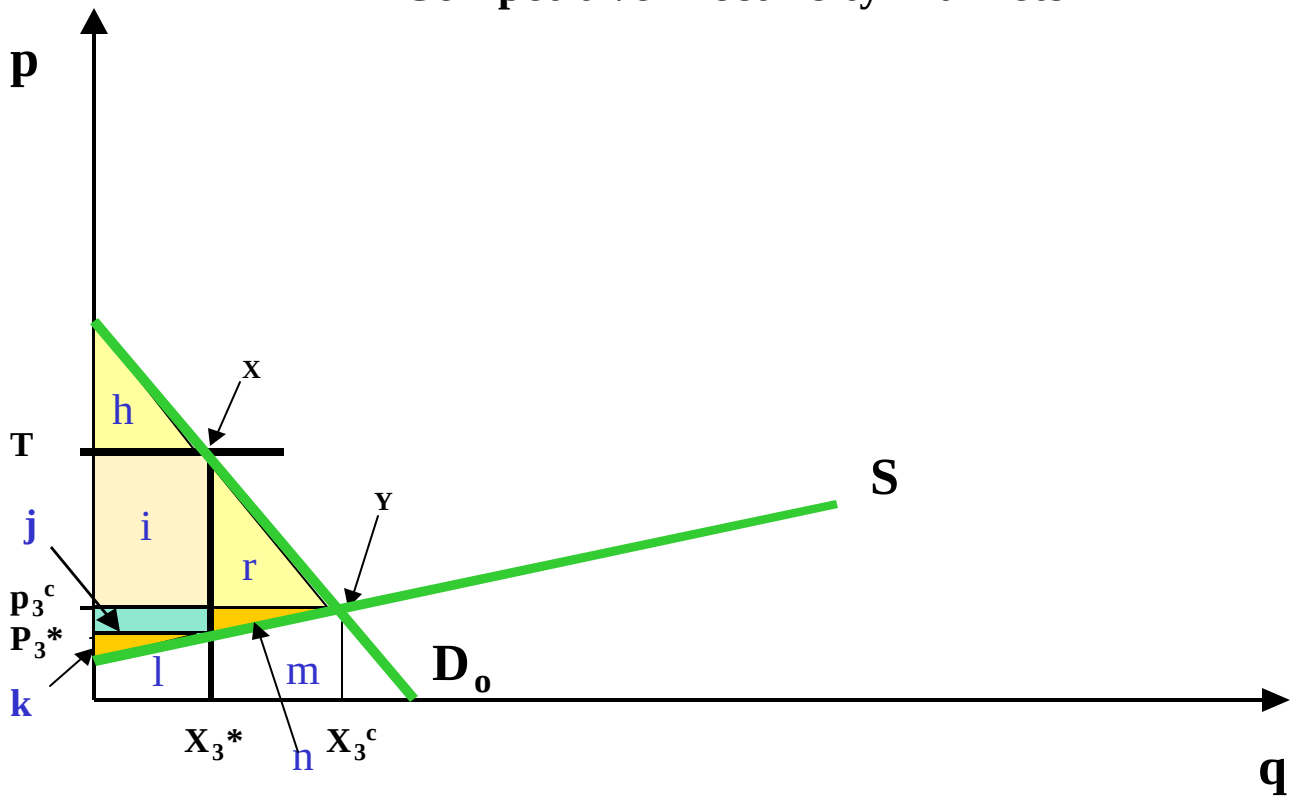


Figure 3. Net Welfare Gain from an Interruptible Load Bidding PRL Program in Competitive Electricity Markets

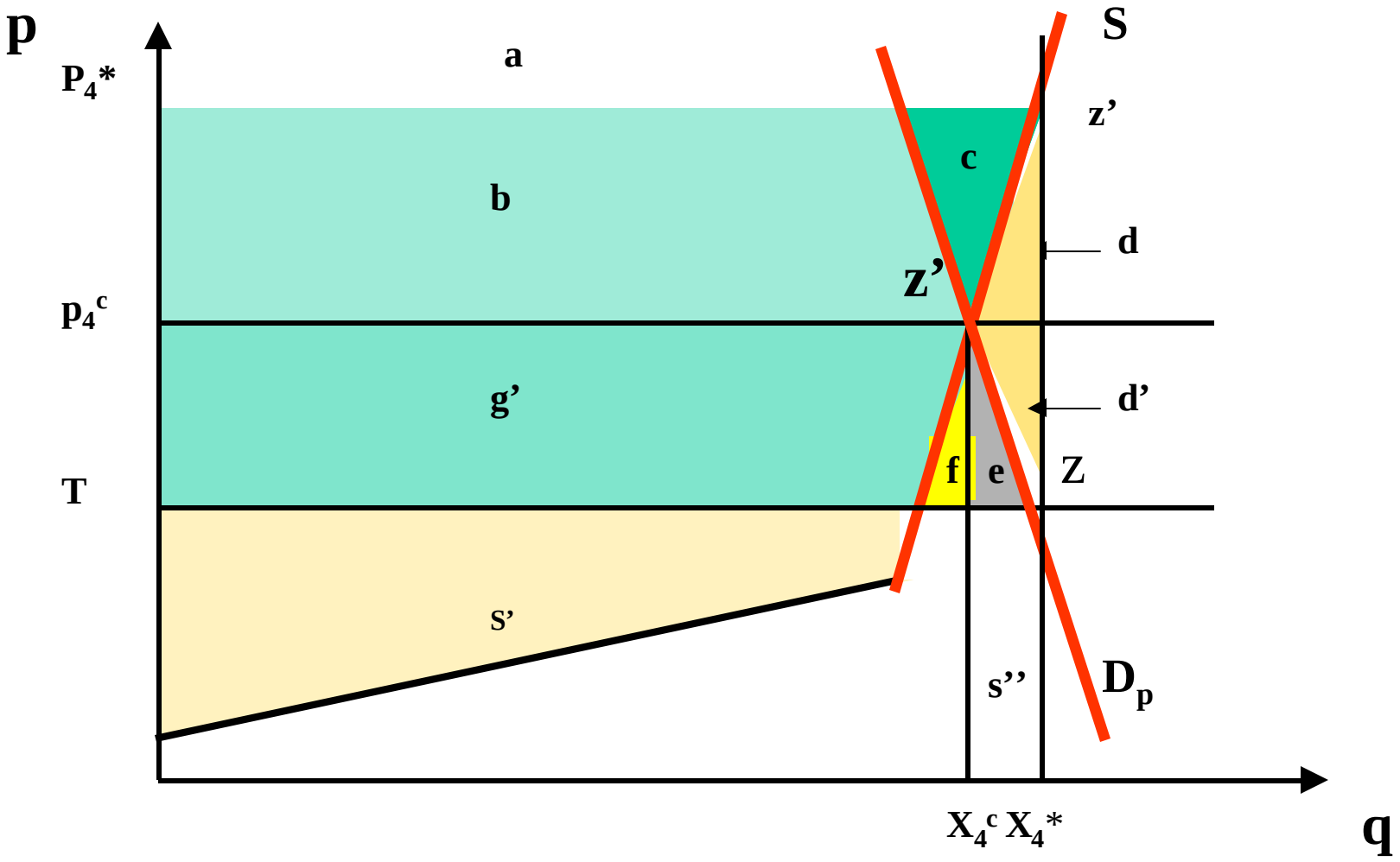


Figure 4. The Relationship Between System Security and System-wide Reserves

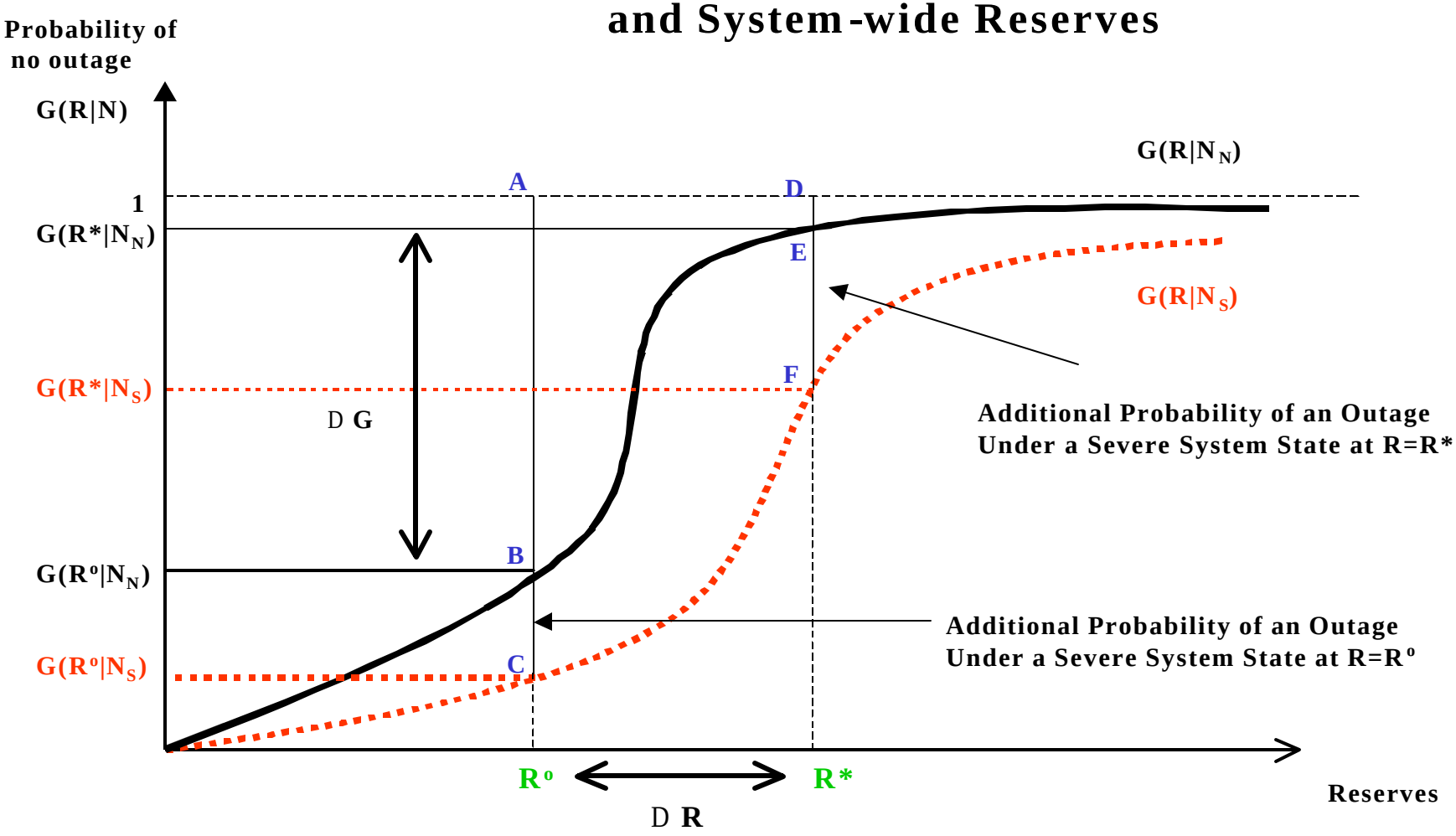


Figure 5. The Relationship Between A Firm's Proportionate Output Loss and System-wide Reserves