LBNL-60132

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Incentives and Rate Designs for Efficiency and Demand Response

Steven Braithwait, Daniel G. Hansen, Laurence D. Kirsch

March 2006

LBNL-60132
Collaborative Report

Incentives and Rate Designs for Efficiency and Demand Response

Steven Braithwait
Christensen Associates Energy Consulting, LLC

Daniel G. Hansen
Christensen Associates Energy Consulting, LLC

Laurence D. Kirsch
Christensen Associates Energy Consulting, LLC

4610 University Avenue, Suite 700
Madison, Wisconsin 53705-2164

March 2006

This work described in this report was coordinated by the Demand Response Research Center and funded by the California Energy Commission, Public Interest Energy Research Program, under Work for Others Contract No. 500-03-026 and by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

PREFACE

The Public Interest Energy Research (PIER) Program supports public interest energy research and development that will help improve the quality of life in California by bringing environmentally safe, affordable, and reliable energy services and products to the marketplace.

The PIER Program, managed by the California Energy Commission (Commission), annually awards up to $62 million to conduct the most promising public interest energy research by partnering with Research, Development, and Demonstration (RD&D) organizations, including individuals, businesses, utilities, and public or private research institutions.

PIER funding efforts are focused on the following six RD&D program areas:

- Buildings End-Use Energy Efficiency
- Industrial/Agricultural/Water End-Use Energy Efficiency
- Renewable Energy
- Environmentally-Preferred Advanced Generation
- Energy-Related Environmental Research
- Energy Systems Integration

What follows is the final report for the Incentives and Rate Design for Energy Efficiency and Demand Response Project, 500-03-026, Task 4.H, conducted by Christensen Associates Energy Consulting, LLC. The report is entitled "Incentives and Rate Designs for Efficiency and Demand Response." This project contributes to the Energy Systems Integration Program.

For more information on the PIER Program, please visit the Commission's Web site at:

http://www.energy.ca.gov/research/index.html or contact the Commission's Publications Unit at 916-654-5200.

Acknowledgements

This work described in this report was coordinated by the Demand Response Research Center and funded by the California Energy Commission (CEC), Public Interest Energy Research (PIER) Program, under Work for Others Contract No. 500-03-026 and by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors wish to thank Roger Levy and Mary Ann Piette for support and valuable suggestions and comments offered in the course of the project.

We would also like to thank the members of our Technical Advisory Committee:

- Barbara Barkovich, Barkovich & Yap, Inc.
- Andrew Bell, Pacific Gas and Electric Company
- Ken Corum, Northwest Power and Conservation Council
- Susan Covino, MADRI - PJM
- David Hungerford, California Energy Commission
- Chris King, eMeter
- Pat McAuliff, California Energy Commission
- Perry Sioshansi, Menlo Energy Economics
- Rick Codina, SMUD
- Russ Garwacki, SCE, and
- Bruce Kaneshiro, California Public Utility Commission,

for their input, suggestions, and comments.

ABSTRACT

This report develops a conceptual framework for designing retail electricity rate structures that provide appropriate incentives for energy efficiency and demand response. The conceptual framework is based upon well-established economic theory of public utility pricing going back at least twenty years, and upon power industry experience of a similar length of history. The emphasis within this document is on the proper application of pricing principles in designing a portfolio of products that will produce the efficient amount of demand response. The report also describes prototype rate designs that illustrate the types of retail rates that provide these incentives. Finally, the report includes a proposed plan for a follow-on Phase II effort that will demonstrate the use of the framework as a tool for long-term research concerning electricity pricing, and will develop, through a utility case study, specific recommended rate structures for use by the California utilities.

TABLE OF CONTENTS

List of Figures

Executive Summary

"Responsive electricity demand" is electric load that changes in response power system conditions, because of either price signals or curtailment calls. The need for responsive demand arises from a combination of: a) the variability of the marginal cost of generating and delivering electricity to consumers; and b) the fixed retail rates that most consumers typically face. The marginal costs of providing customers with power, which are defined as the changes in power system costs that accompany a small increase in demand, vary over time and locations. Marginal costs change over time because electricity is essentially non-storable and because loads change from hour to hour: so different generators with different fuel costs are the marginal power source at different times. Marginal costs change by location because transmission losses and constraints make it more costly to deliver power to some locations than to other locations. Wholesale energy prices generally approximate marginal energy costs. During periods of relative shortage of generating capacity, they can also rise to levels that reflect marginal outage, or capacity costs.

While marginal costs vary over time and location, the retail rates faced by consumers are typically fixed for long time periods. In California, it is primarily large customers who see prices that differ by time-period; and these prices are in the form of time-of-use demand and/or energy charges that vary by time period but are fixed within each time period. Only for a relatively few customers on voluntary critical peak pricing (CPP) or real-time pricing (RTP) programs do prices vary hourly to reflect variations in hourly wholesale costs.

Frequent wide differences between wholesale electricity costs and retail prices suggest extensive foregone opportunities for improvements in economic efficiency. That is, in an important but relatively small number of hours, the marginal cost of producing electricity far exceeds the incremental value that customers receive from consuming it, as reflected in the price they willingly pay. Reductions in usage during these hours would save resource costs (*e.g.*, fuel and capacity costs) far in excess of customers' foregone value of power. However, it is also important to recognize that typical fixed retail prices give consumers no access to the relatively low-cost power that is available in a large fraction of all hours. Increased usage during these periods can produce value to consumers that exceeds the resource cost of generating that power.

Mechanisms for Achieving Responsive Demand

There are two broad types of mechanisms for achieving responsive demand: 1) sending *price* signals to customers, to which customers can respond if they so choose; and 2) sending *quantity* (curtailment) signals to customers, which customers must obey under threat of penalty. For customers, price signals have the advantage of choice: customers can, in effect "buy through" periods of high marginal cost at prices that reflect wholesale market conditions, or decline to accept a payment to reduce consumption.

The most flexible and accurate pricing design is real-time pricing (RTP), in which prices vary hourly and may be announced on the previous day or with only an hour's notice. Because of the potential price risk, most RTP customers will wish to combine hourly pricing with a financial risk management tool that limits their price risk.

Two variations of traditional time-of-use (TOU) pricing provide a dynamic capability to signal periods of unusually high costs. These variations are *critical peak pricing* (CPP) and *day-type TOU* pricing. Under CPP, prices are normally fixed at predetermined levels, but may be increased temporarily, on short notice, to a pre-established critical value whenever wholesale prices rise above or reserve capacity falls below established critical values. Because a utility's expected cost during non-CPP hours is reduced, the prices for those time periods will be lower than under a fixed-price rate. Although CPP has generally been tested in conjunction with an underlying TOU rate, it can just as well be appended to a standard flat or seasonal tariff, with the critical price applying only during specified hours on critical days.

The Cost Basis of Efficient Retail Rates

Efficient retail rates should reflect the marginal costs of power supply. However, they also need to recover regulated utilities' total costs of power supply. A recent trend in retail electricity ratemaking has been the unbundling and separate pricing of electricity services, where the services include the following categories: customer services; distribution services; transmission services; and generation services. The marginal costs that are of most interest from the standpoint of time variation are those of the competitive generation services, which include energy, regulation, and reserves. Because of transmission losses and constraints, these marginal costs may vary by location.

Principles for Efficient Retail Rates

The key to the success of a variable-price rate is in setting the appropriate prices for generation services, including fuel and capacity costs. The appropriate retail price structure will depend in part on the wholesale market design, particularly the services (e.g., energy, capacity) that are included in the market and the ways that those services are priced. For example, if energy, regulation, and reserves are subject to market-based pricing, it may be appropriate to recover generation costs only through retail energy charges. If there is also a capacity market, it may be appropriate to recover capacity costs through retail demand charges.

In the absence of capacity markets, our view is that energy prices (*i.e.*, based on kWh of energy consumption), rather than demand charges (*i.e.*, based on kW of maximum demand), should be used to recover *all* costs of generation services. If energy prices are set appropriately, they will ultimately reflect both the variable operating costs and the capacity costs that consumers are expected to impose on the system, while maintaining a clear and appropriate price signal at the margin. If California institutes a capacity market that has prices analogous to demand charges, then it may be appropriate to recover those costs through retail demand charges that mirror the capacity charges, although those costs could still be recovered through appropriately designed energy prices that reflect the probability of incurring capacity charges.

The report provides a mathematical description of the appropriate method for calculating any fixed or time-varying retail energy rate. In fact, the key to offering consumers a choice among rate options is to use the same fundamental algorithm for each rate, accounting for differences in price risk implied by the features of the rate. The pricing algorithm calculates the expected cost of serving the relevant customer load, taking into account the risk associated with uncertainty of both loads and prices over the time periods to which the rate applies.

Customers have different preferences for price certainty. Some customers—particularly those who are most able to respond to price changes—are willing to face prices that can change with little notice and that may vary substantially from hour to hour and day to day. Other customers—particularly those who are unable to shift loads among time periods—prefer prices that are stable and announced well in advance; and they will be willing to pay a premium to avoid price variations and price uncertainty. Offering a menu of retail rates that involve different degrees of price risk can lead to the efficient amount of demand response that represents all consumers' combined willingness to curtail load in response to high market prices or to pay a premium for the certainty that power will be available at a guaranteed price.

Conclusions

The implications of the conceptual framework developed in this report are straightforward. First, retail rates that provide appropriate incentives for demand response are those whose prices reflect the time-varying and location-varying nature of the marginal costs of generating and delivering electricity to the relevant group of consumers. Second, a variety of time-varying rate structures may be designed, each of which reflects expected marginal costs with different degrees of accuracy, price guarantees, and notice of price changes. Third, offering a limited menu of such rate structures, properly designed to provide consistent risk premiums, will lead to the appropriate amount of demand response that consumers are willing to provide. Finally, while the implications of the framework regarding retail pricing are clear, existing rate regulation practices may hinder the efficient design of both default and time-varying retail rates. The effect of these practices should be explored and addressed in follow-on research in Phase II.

Phase II Plan

The proposed Phase II project has three primary objectives: 1) to assess the existing status of retail rates in California, including recently proposed CPP and RTP rates, relative to criteria consistent with the framework defined in this report; 2) to work with project stakeholders to assess the existing regulatory barriers to adopting retail rates that provide appropriate incentives for efficiency and demand response; and 3) to work with a case study utility to develop an "ideal" set of default and optional rates that provide such incentives.

To achieve these objectives, Phase II will involve several sets of activities, all of which will be conducted in close collaboration with the technical advisory group, the utilities, the CEC, and the CPUC. A key element of the Phase II research will be the involvement of at least one case study utility, which will be solicited from the three California IOUs.

The primary analysis activity of the Phase II research will involve applying the framework developed in Phase I, as extended in light of ideas and information developed in Phase II, to develop, quantify, and evaluate recommended menus of standard and voluntary retail rate structures.

1.0 Introduction

Well-functioning wholesale electric power markets require some degree of retail demand response to power market conditions. Such response is most critical during power shortage conditions, when load reductions can improve power system reliability and can help avoid costly fuel expenditures. By reducing peak loads, demand response can also facilitate reductions in capital expenditures on generating capacity, and can help reduce the ability of suppliers to exercise market power. In off-peak periods, demand response can provide additional benefits to customers by giving them cheap power when power is plentiful.

By contrast, however, most electricity is sold to consumers at fixed retail prices that bear no relationship to current power system conditions. This traditional pricing approach provides little or no incentives for consumers to respond to power system conditions. The consequence for electricity markets has been characterized as being similar to driving a car without shock absorbers: the potential jolts from unexpected potholes (representing occasional conditions of low generating reserve capacity with their accompanying price spikes) create a need for extra spending to guarantee smooth roads (representing extra reserve capacity to meet consumers' non-responsive demands). That is, the lack of responsive demand can lead to higher generation capacity and higher overall energy costs than if demand were responsive.

1.1. Objectives of This Report

This report has been developed at the behest of the Demand Response Research Center (DRRC), which is funded by the California Energy Commission Public Interest Energy Research Program (PIER) and managed by the Lawrence Berkeley National Laboratory. This report is also partly motivated by research needs identified by the California Public Utilities Commission.[1]

To meet the research needs of the foregoing organizations and of electricity consumers in California, this report develops a conceptual framework for designing retail electricity rate structures that provide appropriate incentives for demand response, and for calculating the economic benefits from such rates. In addition to redesigning retail rates to better reflect wholesale costs, and thus induce demand response when it is most needed, industry stakeholders in California are interested in improving consumers' overall energy efficiency. There is thus interest in retail rate designs that provide consistent appropriate incentives for both demand response and energy efficiency.

The objective of this research project is to develop the following two key products:

1. *A conceptual framework* for integrating and improving the effectiveness of incentives used to support efficiency and demand response, and
2. *Prototype rate designs* that illustrate the application of the framework for residential, commercial, and industrial customers.

The conceptual framework facilitates analysis of tariffs that provide incentives for efficiency and demand response, and the prototype rate designs illustrate the types of designs that provide these incentives. Our proposed plan for a follow-on Phase 2 effort will demonstrate the use of the framework as a tool for long-term research concerning electricity pricing, and will

[1] See, for example, California Public Utilities Commission [2005].

develop specific recommended rate structures for use by the California utilities. The framework and proposed research plan consider the relationship among the economic incentives that those rates provide for customers to take energy efficiency actions, the policy support for providing new incentives for customers to take demand response actions, and the current retail rate designs in California.

Our conceptual framework is based upon well-established economic theory of public utility pricing going back at least twenty years, and upon power industry experience of a similar length of history. The emphasis within this document is on the proper application of pricing principles in designing a portfolio of products that will produce the efficient amount of demand response.

1.2. Organization of This Report

This report begins, in Section 2, with a discussion of the reasons for and benefits of demand response. Section 3 describes the price and non-price mechanisms for achieving demand response, while Section 4 explains how efficient retail rates depend upon costs. Sections 5 and 6 specify the mathematics and other considerations for quantifying the efficient prices for price and non-price demand response programs. Finally, Section 7 provides a plan for Phase II of this project.

The main body of the report is followed by three appendices. Appendix A reviews the history of the theory of public utility pricing. Appendix B describes customers' bill and incentives under hedged real-time pricing. Finally, Appendix C discusses the consistency of retail pricing incentives for efficiency and demand response.

2.0 The Need for Demand Response

The need for responsive electricity demand[2] arises from a combination of: a) the variability of the marginal cost of generating and delivering electricity to consumers; and b) the fixed retail rates that most consumers typically face. Consequently, this section first reviews the nature of and recent historical patterns of electricity marginal costs in California.[3] It then reviews the effects of a lack of responsive demand, illustrates the opportunities for demand response, and provides a framework for assessing the economic benefits that may be achieved by modifying existing retail tariffs to better reflect marginal costs.

2.1. Marginal costs

The marginal costs of providing customers with power vary over time and locations. "Marginal costs" are defined as the change in power system costs that accompanies a small (e.g., 1 MW) increase in demand (or "load"). Marginal costs change over time because electricity is essentially non-storable and because loads change from hour to hour, and indeed, from minute to minute: so different generators with different fuel costs will be the marginal power source at different times. Marginal costs change by location because transmission losses and constraints make it more costly to deliver power to some locations than to other locations.[4]

In principle, the marginal costs associated with changes in load levels are composed of three broad categories of costs. First is marginal *energy* costs (also called "marginal operating costs"), which are primarily the marginal fuel and labor costs of generating power. Second is marginal *outage* costs (also called "marginal reliability costs" or "marginal capacity costs"), which are the expected costs of the risk that a load increase will increase the chance of a generation-related power shortage. Marginal outage costs measure the reliability benefit that is due to a load reduction. Third is marginal *externality* costs, which are costs imposed upon parties other than electric power market participants. Prominent among these latter costs are environmental and national security costs.

Wholesale energy prices generally approximate marginal energy costs. During periods of relative shortage of generating capacity, they can also rise to levels that reflect marginal outage costs.

Figure 1 shows the distribution of hourly wholesale energy prices in Southern California (region SP15) in the years 2002 through 2005. The figure shows hourly wholesale prices for the summer months (June through September) of each year, arrayed from high to low. The graph show similar patterns of wholesale prices in each of the four years, with relatively low prices in approximately a third of the hours, relatively similar prices in approximately half of the hours,

---

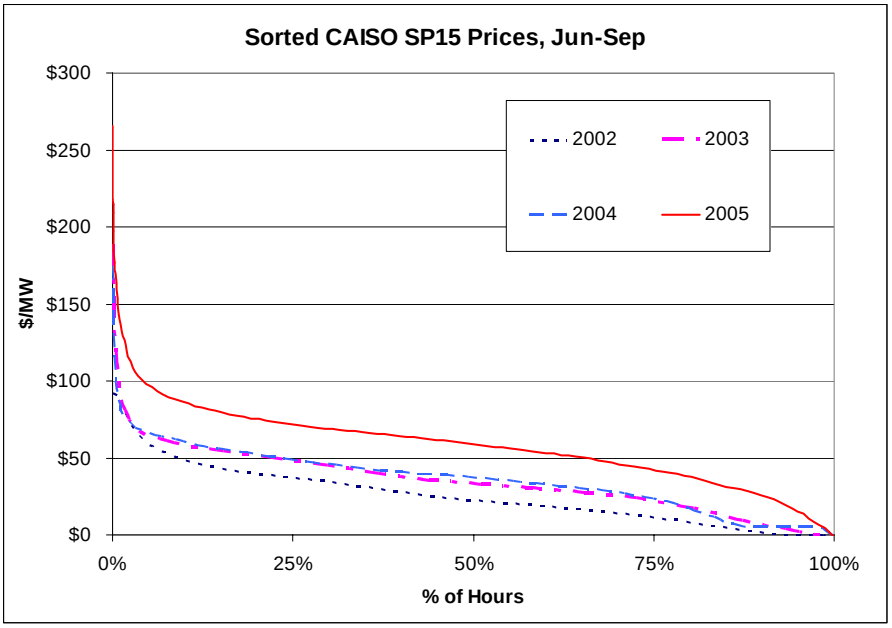
[2] The term "responsive demand" is used in this report to refer to the condition under which consumers face retail prices that reflect variations in wholesale power costs, and modify their usage patterns in response to those prices.

[3] Appendix A provides an historical review of the theory of public utility pricing and the role of marginal costs.

[4] See Section 4.1 for a more complete description of how power systems operate and how marginal costs change over time and space.

and relatively high prices in the highest 10% or so of hours. The generally higher prices in all hours in 2005 reflect higher natural gas prices in that year.

Figure 1
Frequency Distribution of Southern California Wholesale Energy Prices
Summers of 2002-2005



2.2. Effect of Non-Responsive Demand

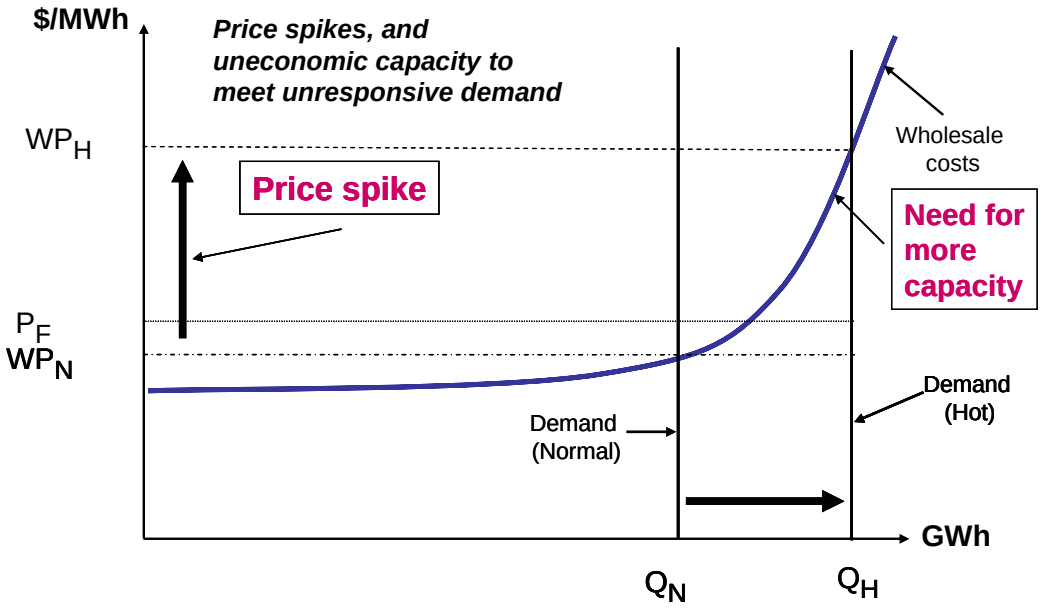
Although marginal costs vary over time and location, retail rates faced by consumers are typically fixed for reasonably long time periods. This has been the case because traditional utility rate design has focused largely on recovery of allowed costs and on methods for allocating those costs fairly across various customer types. The economic efficiency of the resulting price structures, in the sense of establishing prices that reflect utilities' time-varying marginal costs, has typically been given a low priority. As a result, while marginal costs and wholesale power prices vary hourly, retail power prices for most customers are either fixed over time or vary seasonally at most. In California, for example, the retail prices paid by some customers are higher during the summer months than in non-summer months. Typically, however, only large customers see prices that differ by time-period; and these prices are in the form of time-of-use demand and/or energy charges that vary by time period but are the same within each time period. Only for a relatively few customers on critical peak pricing (CPP) or real-time pricing (RTP) programs do prices vary hourly to reflect variations in hourly wholesale costs.

Figure 2 illustrates the effect of the lack of responsive demand in the presence of varying wholesale costs. It features a simplified electricity supply curve for a particular afternoon hour. The supply curve is upward-sloping because generators with lower incremental energy costs provide output first, while generators with higher incremental energy costs provide power only when load become high. As is typical of the supply curves of many power systems, this supply

curve rises steeply at levels of output that approach overall system capacity. The figure also shows two different demand levels, Q_N and Q_H, corresponding to Normal and Hot weather conditions. These demand curves are shown as vertical lines because consumers face a fixed retail price (P_F) and therefore have no incentive to modify their usage levels under different power system conditions. The demand levels and marginal cost curve combine to produce wholesale energy prices under the two scenarios, where the prices reflect conditions in, for example, a day-ahead energy market.

As shown in the figure, at normal load levels, the wholesale price (WP_N) takes on a relatively low value. At the higher level of demand, the wholesale price rises to a much higher level (WP_H). This reflects the historical experience in certain regions, as demand on occasion has soared due to extreme weather conditions, and/or unexpected outages have constrained capacity, causing wholesale prices to spike to levels reaching as high as $8,000 per MWh ($8.00 per kWh).[5] Such values reflect not only the high operating cost of the generating units that are the last ones called to run, but also the market's valuation of maintaining reliability and avoiding costly outages. Occurrence of such price spikes signal the need for additional generating capacity to meet the level of non-responsive demand with adequate reliability. Implementing retail rates that reflect varying wholesale costs produces responsive demand that can avoid the need for such additional capacity.

Figure 2
Effect of No Responsive Demand



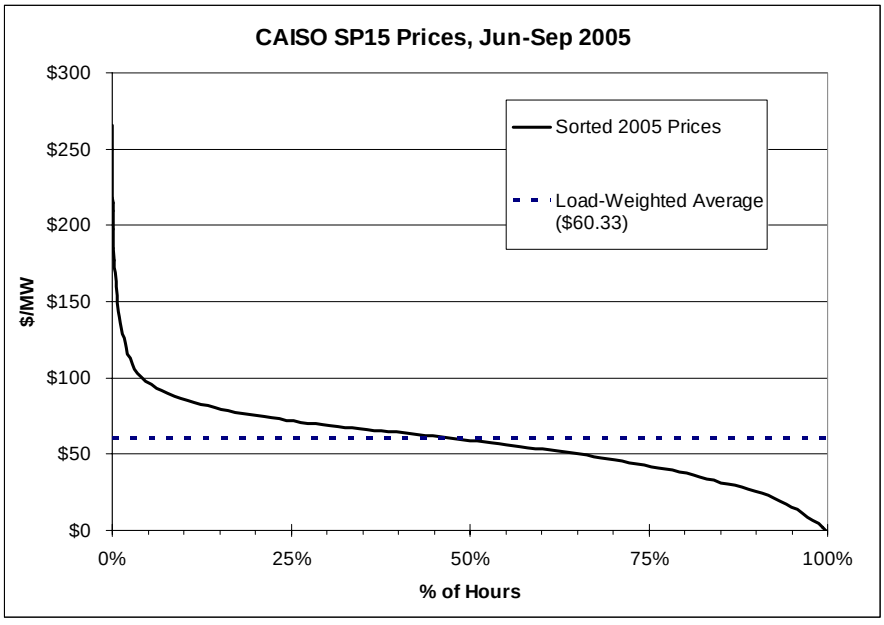
[5] There is some debate over whether extremely high prices have been caused by true scarcity or by artificial scarcity due to the exercise of market power by generators.

2.3. Opportunities for Demand Response

Figure 3 combines the distribution of Southern California hourly wholesale prices for 2005 with a horizontal line showing the CAISO system load-weighted average of wholesale prices during the summer (June-September) season. The latter value represents a flat seasonal energy price that would recover the same amount of revenue as if customers were charged hourly wholesale prices. The figure demonstrates the disconnection that can exist between *varying* wholesale energy costs and *fixed* retail prices. Some important features of the values shown in the figure are that in approximately 25% of the hours of the summer, wholesale electricity costs were at least a third less than the load-weighted average (sometimes much less), while in approximately 5% of the hours, electricity costs were more than twice as high as the load-weighted average.

Such frequent wide differences between wholesale electricity costs and retail prices suggest extensive foregone opportunities for improvements in economic efficiency. That is, in an important but relatively small number of hours, the marginal cost of producing electricity far exceeds the incremental value that customers receive from consuming it, as reflected in the price they willingly pay. Reductions in usage during these hours would save resource costs far in excess of customers' foregone value of power. However, it is also important to recognize that typical fixed retail prices give consumers no access to the relatively low-cost power that is available in a large fraction of all hours. Increased usage during these periods could produce value to consumers that exceeds the resource cost of generating that power.[6]

Figure 3
Opportunities for Increased Economic Efficiency—Differences Between Marginal and Average Electricity Costs



[6] To the extent that externality costs, such as the environmental costs of power plant emissions are not included in generators' costs, and thus market prices, then those prices do not reflect all resource costs.

Increasing the amount of responsive demand is generally viewed as producing a number of potential benefits, including:

- Lower overall energy costs due to avoiding high operating costs,
- A reduced amount of required generation capacity than that needed to meet non-responsive demand,
- Reduced market power on the part of generators during periods of low reserve capacity,
- Reduced wholesale price spikes, and
- Improved overall market efficiency.

However, proposals to increase the amount of responsive demand through retail rate re-design have run into a number of barriers, including the following:

- The cost of metering equipment needed to support time-varying prices,
- Concern about the complexity of time-varying rates,
- Uncertainty about appropriate rate design,
- Uncertainty about how consumers will respond,
- Concern about bill impacts on certain consumers, and
- Concern about revenue impacts on utilities.

Many policy statements about demand response programs state the objective as maximizing the *amount* of demand response. This goal is likely to produce a program that overpays for demand response. An alternative goal is to maximize social welfare (or net benefits), which can be restated as seeking to obtain the *efficient* amount of demand response. Maximization of social welfare assures that Californians—electricity customers and citizens in general—obtain the greatest possible net benefits from electricity, considering all of the benefits and all of the costs and concerns listed above.

2.4. Economic Benefits of More Efficient Retail Rate Design

This section provides a framework for measuring the economic benefits that may be achieved by modifying retail service offerings so that they better reflect power system conditions and give consumers incentives to respond to those conditions. In general, the net benefit of improving the efficiency of retail rates will equal the change in the value of electricity consumed induced by the new rates minus the change in the cost of electricity provided:

$$\text{Net Benefit} = \Delta\text{Value} - \Delta\text{Cost} \quad (1)$$

where "cost" can be broadly interpreted to include generator operating costs, reliability risks, environmental impacts, and so on. In cases wherein the new rate design induces a *load reduction*, the change in value will be negative because customers will consume less; but if the program is efficient, the change in costs will be even more negative (that is, cost savings will be relatively large) so that the net benefit is positive. In a case wherein the new rate induces a *load increase*, the change in value will be positive because customers will consume more; but if the

program is efficient, the change in costs will less positive so that the net benefit will (again) be positive.

One interpretation of the economic effect of responsive demand is that the load reductions by price-responsive consumers allow them to effectively serve as virtual generators, substituting a load reduction for an equivalent amount of peaking generation capacity or high-cost power purchases that would otherwise be needed to meet high demand that was *not* price-responsive. To the extent that the combination of consumers' cost of curtailing load and the administrative costs of enabling responsive demand is less than the cost of additional generating capacity on the supply side, then such load reductions will reduce overall energy costs. Consumers will reveal their curtailment costs through their willingness to participate in dynamic pricing or demand response programs, and through the MW amounts by which they are willing to curtail load at different price levels.

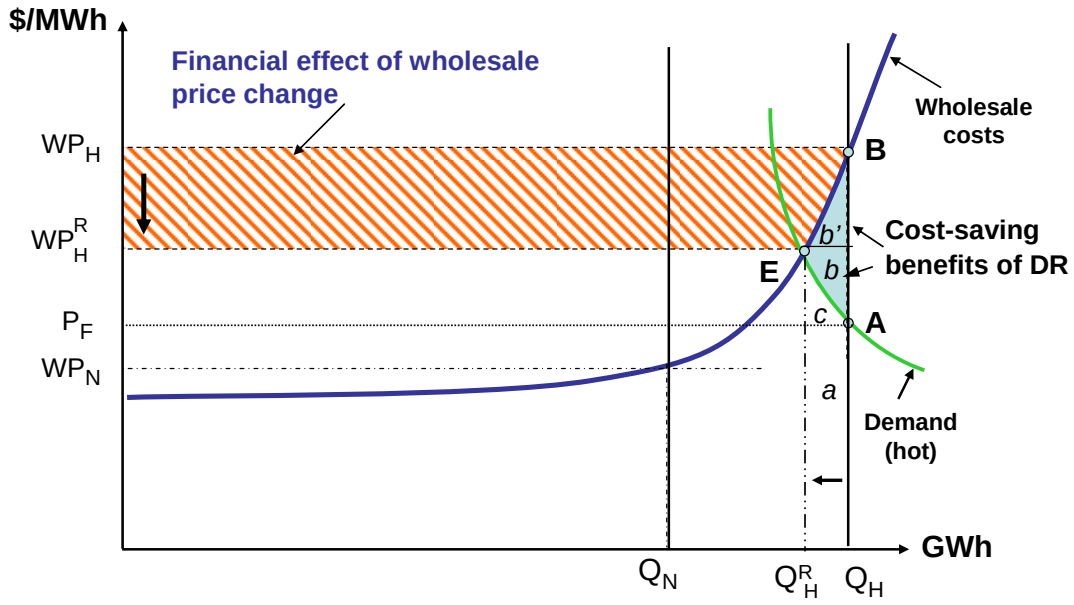
2.4.1. A demand reduction case

Figure 4 illustrates a case wherein supplies are tight relative to load. Like Figure 2, Figure 4 shows conditions in a representative hour in a day-ahead energy market. The figure contains three alternative market demand curves, which specify the amount of electricity demanded by consumers at various price levels. Two are vertical lines representing alternative load levels under normal (N) and hot (H) summer conditions when consumers face fixed retail prices (P_F), and therefore have no incentive to change their usage patterns when wholesale market prices rise. The third demand curve, labeled Demand (hot), slopes downward to reflect the responsive demand that consumers reveal when they face time-varying retail prices that reflect wholesale costs. That is, when consumers face higher retail prices, they reduce load; and when they face lower retail prices, they increase load.

Key elements of Figure 4 are as follows:

- On a hot summer day *without responsive demand*, demand is Q_H, leading to wholesale prices of WP_H.
- The marginal *cost* of producing the last unit of power to meet demand under the no responsive demand scenario is given by the vertical distance from the horizontal axis to point *B* on the supply curve. The incremental *value* to consumers of that last increment of demand is shown by the height to point *A* on the aggregate demand curve at the fixed retail price P_F. That is, when the price is P_F, consumers use electricity up to the point where the incremental value they obtain is just equal to that price.
- If some consumers face dynamic prices that reflect wholesale costs, or can offer load curtailments into the wholesale market through a demand response program, then aggregate demand is shown by the revealed sloping demand curve. In that case, the wholesale market clears at point *E*, where price is WP_H^R and the total quantity demanded at that price is Q_H^R.

Figure 4
Economic Benefits of Responsive Demand
(high-cost hour)



Two primary effects of responsive demand may be seen in the figure. First, the amount of electricity consumed falls under scarcity conditions, thus saving the system the high cost of generating the electricity that is not consumed. Second, the severity of the wholesale price spike is lessened. The first effect represents a change in the use and allocation of economic resources, which is the focus of economic efficiency. The second effect is financial, representing wealth transfers among market participants (*i.e.*, consumers and generators), but no net change in the efficient use of resources.

The overall net increase in economic benefits due to demand response is represented by the shaded area (b + b′) in Figure 4. Consistent with equation 1, this net increase consists of cost savings less foregone value of consumption. The cost savings are represented by the area under the supply curve for the amount of the load curtailment, ($a + c + b + b'$). The foregone value of consumption illustrated by the area under the demand curve for the amount of the curtailment, ($a + c$). The net benefit is the cost savings minus the foregone value, ($a + c + b + b'$) – ($a + c$) = (b + b′).

Depending on how dynamic pricing or demand response programs are designed, the overall net cost savings ($b + b'$) are available to be *shared* between the responsive consumers who curtail load, their energy suppliers whose energy supply costs are reduced, and potentially all consumers due to lower overall power costs. These allocations of economic net benefits are described below.

2.4.1.1. Allocation of economic net benefits

To explain how the net benefits of equation 1 are divided among different parties, we note that, as a purely accounting matter, the change in generators' revenues must equal the change in consumers' bills. This is the same as saying that there is no difference between the change in revenues and the change in bills.

$$\Delta\text{Revenue} - \Delta\text{Bills} = 0 \quad (2)$$

We also note that the change in costs can be broken into three components: a change in generation costs; a change in the expected reliability of service; and a change in external costs (like environmental costs).

$$\Delta\text{Cost} = \Delta\text{Generation Cost} + \Delta\text{Reliability Cost} + \Delta\text{External Cost} \quad (3)$$

Using equations 2 and 3, equation 1 can then be expanded as follows:

$$\begin{aligned}\text{Net Benefit} &= \Delta\text{Value} - \Delta\text{Cost} \quad (4)\\ &= \Delta\text{Value} - (\Delta\text{Generation Cost} + \Delta\text{Reliability Cost} + \Delta\text{External Cost})\\ &\quad + (\Delta\text{Revenue} - \Delta\text{Bills})\\ &= (\Delta\text{Value} - \Delta\text{Bills} - \Delta\text{Reliability Cost})\\ &\quad + (\Delta\text{Revenue} - \Delta\text{Generation Cost})\\ &\quad - \Delta\text{External Cost}\end{aligned}$$

In equation 4, the first equality is merely a restatement of equation 1. The second equality inserts into equation 4 the cost breakdown of equation 3 and a zero in the form of the difference between the change in revenues and the change in bills. The third equality rearranges the terms of the second equality.

The terms in the third equality of equation 4 are particularly interesting because they present a breakdown of benefits and costs between consumers, energy service providers, and non-market participants. The first line says that consumers' net benefit equals the change in the value of their consumption less the change in their bills less the change in expected reliability costs. For a peak period load reduction, all the terms in that first line will be negative, so that consumers will lose value; but they will benefit from bill savings and improved reliability, and thus achieve net gains that can make participation attractive. Similarly, the second line says that the net benefit to energy service providers equals the change in their revenues minus the change in their costs. To the extent that their cost savings exceed foregone revenue from customers' bill savings, they also can achieve net gains.[7] The third line says that non-participants are affected by a change in external costs: for example, if a load reduction reduces environmental costs, the cost change will be negative and the effect on non-participants will be positive.

[7] Alternatively, an energy service provider that has available capacity or contracted power during a period of high wholesale prices can benefit from selling, into the wholesale market, power that has been freed-up by demand response.

2.4.1.2. Financial effects of demand response

Some confusion and controversy exist in discussions of the benefits (or economic value) of the shock-absorbing feature of responsive demand, such as the benefits of the reduction in wholesale price spikes shown in Figure 4. Proponents of demand response often cite the financial effects of wholesale price spike reductions as evidence of the large potential benefit of demand response programs. However, there are several problems with this interpretation. First, the bill and revenue effects resulting from *price* changes do not measure changes in real economic resource cost or value. Instead, changes in real economic values require changes in the *quantity* of demand.

Second, in the short run, wholesale price changes have little financial effect on most market participants because nearly all of them manage their price risk by either owning generation or entering financial contracts to buy and sell blocks of energy at fixed prices. As a result, these parties are for the most part *not directly hurt financially by temporary wholesale price spikes, nor helped by reductions in the price spikes*. The financial impact shown in the figure simply represents one of thousands of offsetting hourly financial adjustments that are implied by standard forward contracts at fixed prices in the presence of variable spot prices. Over the period of the contract, such offsetting adjustments between buyers and sellers of those contracts can be expected to balance out over hours of high and low spot prices.

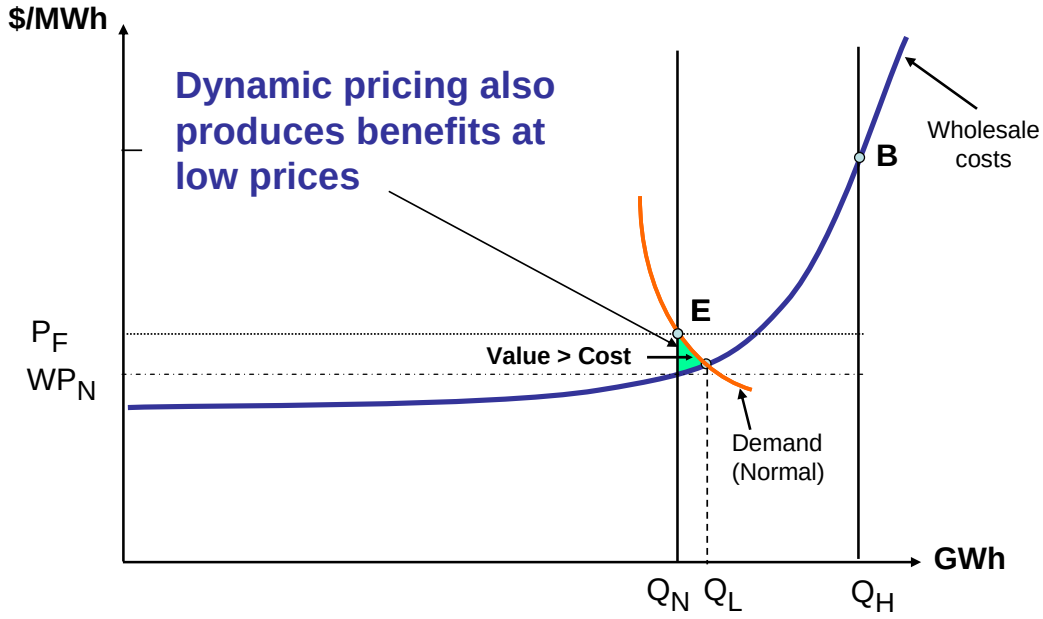
Third, in the long term, wholesale prices reflect the capacity investment decisions made in response to historical prices. Thus, while responsive demand may indeed serve to hold down short-term price spikes, *conditional on the existing generating resources*, generation firms will take such reductions into account when making investment decisions for the future. As a result, short-term price reductions cannot be assumed to hold indefinitely into the future, because the amount of capacity likely to be installed under a scenario *with* responsive demand is likely to be less than *without* responsive demand. As noted above, the long-run effect of the savings from avoiding the cost of capacity that does not have to be built represents one of the benefits of responsive demand.

2.4.2. A demand increase case

Dynamic pricing can also provide economic benefits during periods of low wholesale costs, as consumers are able to use additional electricity for relatively low-valued applications that would not be justified at the higher fixed retail price. Such effects are shown in Figure 5. At the fixed price P_F, consumption under normal conditions is Q_N. With responsive demand, however, consumers will respond to wholesale prices lower than P_F by raising consumption to Q_L, at which output level the wholesale price will be WP_N. This creates net benefits equal to the little shaded triangle under point E.

As suggested by Figure 5, the magnitude of the net benefit in a given hour is likely to be smaller than the potential cost-saving benefits during a high-cost hour. The hourly net benefit is likely to be smaller in the low-cost case because of the relatively smaller differences between the fixed price and varying costs in low-cost periods. However, there are many more low-cost than high-cost hours, implying that the total potential gain from lower prices in low-cost hours may be substantial.

Figure 5
Economic Benefits of Responsive Demand
(low-cost hours)



In practice, the changes in benefits and costs illustrated in Figures 4 and 5 may be quantified, for all hours in a relevant period, using analytical tools and software to represent the customer demand and market supply curves shown in the figures. Given baseline information on hourly wholesale costs, standard retail rates, customer loads, and price responsiveness, such tools may be used to calculate customers' load changes in response to dynamic prices. These tools can also be used to calculate the changes in customer bills and wholesale costs that are implied by those load changes. Such tools can thus assist in the design and evaluation of efficient retail rate structures intended to provide incentives for efficiency and demand response.

In one study of this type, Braithwait and Armstrong [2004] estimated the potential impact of RTP in California using marginal cost and customer load data for California, and customer price responsiveness estimates derived from evaluations of Georgia Power Company's RTP program. Results from this study have been used in some of the California utilities' business case plans for Automated Metering Infrastructure. In another study, Borenstein [2005] assessed the potential benefits of RTP for a generic power system under alternative assumptions about cost conditions and RTP market share.

3.0 Mechanisms for Achieving Responsive Demand

There are two broad types of mechanisms for achieving responsive demand: 1) sending *price* signals to customers, to which customers can respond if they so choose; and 2) sending *quantity* (curtailment) signals to customers, which customers must obey under threat of penalty. For customers, price signals have the advantage of choice: customers can, in effect, "buy through" at prices that reflect wholesale market conditions, or decline to accept a payment to reduce consumption. For power system operators, quantity signals have the advantage of greater control: operators can control customer loads more directly than with price signals.[8]

Price signals may be sent directly through retail rates, or indirectly through *demand response programs*, in which consumers receive a payment for reducing load relative to a baseline level of consumption, where the payment reflects the value of the load reduction in the wholesale market. The payment may result from an offer by the consumer in a demand-bidding program, or from the consumer's acceptance of an offer from his energy service provider (ESP). Similarly, quantity signals may be sent from a utility through a traditional load management program, or from an ESP or independent system operator through certain types of demand response programs.

3.1. Pricing Mechanisms

The frequent differences between the cost of generating electricity and the fixed price that most consumers pay may be narrowed by more refined pricing designs in which retail prices vary in some degree to reflect time-varying costs. Two fundamentally different categories are: 1) *time-of-use* (TOU) rates, in which prices generally vary by season and time of day, but are fixed for a relatively long period of time; and 2) *dynamic pricing*, in which at least some prices may be changed at reasonably short notice (*e.g.*, a day ahead or an hour ahead of when they go into effect). These pricing designs are referred to generally in the Energy Policy Act of 2005 as *time-based pricing*. A few basic examples are described here. The range of efficient pricing designs is discussed in greater detail in Section 5.

TOU prices are held constant within each TOU period. For example, a *summer peak* price might apply to specified summer weekday afternoon hours that tend to correspond to the hours of highest costs; and the same price would apply to all of these hours throughout the summer. A simple TOU rate structure might have only a peak and an off-peak price, both of which apply to their respective TOU periods throughout the whole year. A more refined design would have prices that vary by season, and might also have a shoulder (or mid-peak) price that applies to blocks of hours on either side of the peak period to reflect hours of intermediate cost. A

[8] For a discussion of the relative merits of price and quantity signals, see Weitzman [1974]. Weitzman says that price signals are better than quantity signals when distortions in quantities away from efficient levels have small efficiency implications, but that quantity signals are better than price signals when a quantity error (like electricity demand exceeding electricity supply) would have dire implications. As applied to electricity, Weitzman's argument implies that quantity signals are superior to price signals under conditions of power system stress because small errors in quantity can lead to large costs (e.g., power system outage), while small errors in price are of relatively little importance. Hence, during periods of potential power scarcity, it is more important to get quantities right (even if prices are wrong) than to get prices right (while getting quantities wrong).

fundamental deficiency of TOU prices is that they do not differentiate days on which wholesale prices are relatively low from the occurrence of a day with unusually high energy costs, on which consumer demand response is most valuable.

Two variations on TOU pricing that provide a dynamic capability to signal periods of unusually high costs are *critical peak pricing* (CPP) and *day-type TOU* pricing. Under CPP, prices are normally fixed at predetermined levels, but may be increased temporarily, on short notice, to a pre-established critical value whenever wholesale prices rise above or reserve capacity falls below established critical values. Depending on the design, consumers receive notice on the previous day or a few hours before a critical event. Although CPP has generally been tested in conjunction with an underlying TOU rate, it can just as well be appended to a standard flat or seasonal tariff, with the critical price applying only during specified hours on critical days.

Under day-type TOU pricing, two or three sets of TOU prices are pre-established (*e.g.*, low, medium, and high) to reflect average wholesale energy cost patterns on typical day types. The price day-type is typically announced in the afternoon of the preceding day. This type of price structure is offered to residential consumers by Electricité de France (EdF) as the *Tempo* rate.[9]

The most flexible and accurate pricing design is real-time pricing (RTP), in which prices vary hourly and may be announced on the previous day or with only an hour's notice.[10] Because of the potential price risk, most RTP customers will wish to combine hourly pricing with a financial risk management contract that limits their price risk.

Section 5 describes rate design strategies for achieving efficient retail prices subject to revenue recovery of all allowed costs.

3.2. Curtailable Service Programs

In the absence of efficient retail prices, utilities have traditionally offered some form of *load management* programs, such as direct load control of residential air conditioners and water heaters, and curtailable service programs for large industrial customers. In both cases, rate discounts are typically offered to customers in return for agreements that those customers will curtail usage when requested by the utility, typically only under emergency conditions when reliability is at risk. In areas where retail competition has been allowed and regional transmission organizations (RTOs) or independent system operators (ISO) have been established, load management programs have largely been replaced or re-cast to coordinate with ISO-sponsored emergency- or capacity-related demand response programs. In these programs, consumers receive a capacity credit payment in return for offering to reduce load at short notice under conditions of system reliability risk.

Many traditional curtailable service programs have merely been vehicles for giving discounts to large industrial customers, even though interruptions hardly ever occur. In contrast, the California crisis of 2000-2001 was one instance in which curtailable customers were actually required to sustain the maximum number of allowable curtailments.

---

[9] See Aubin *et al* [1995].

[10] RTP prices in some areas have been established as *ex post* prices that reflect the actual prices produced in a real-time energy market, typically in cases where no day-ahead market exists. This design gives consumers no advance notice of prices, which makes demand response difficult.

Section 6 of this report describes the development of efficient, market-based curtailable service programs that are consistent with efficient retail price structures.

4.0 The Cost Basis of Efficient Retail Rates

Efficient retail rates should reflect the marginal costs of power supply. However, they also need to recover regulated utilities' total costs of power supply.

This section first describes how electricity costs may be unbundled into the costs of providing certain fundamental services. It then describes the nature of the marginal costs of power supply, and then the types of information on marginal costs that will be produced by the planned California wholesale power market design and available for use in designing efficient retail rates.

4.1. Cost Unbundling

A recent trend in retail electricity ratemaking has been the unbundling and separate pricing of electricity services. The trend to unbundle costs has been driven largely by industry restructuring, including the move to retail competition in some states. Distribution utilities whose customers may purchase energy from an alternative supplier must have a way to identify the non-energy costs to serve those customers and establish rates to recover those costs.

Utilities in states that have undergone restructuring, including California, have typically unbundled their costs into the following four categories of basic services:

- *Customer services*, which primarily involve communication with customers (as through billing). The costs of these services generally depend upon the number of customers and the passage of time, and do not vary with customers' energy consumption.
- *Distribution services*, which are services associated with distribution facilities. In any subsystem of the distribution system, these costs depend on the number, location, and density of customers, and on peak loads.
- *Transmission services*, which are services associated with transmission facilities. The costs of these services can be divided into two broad categories: transmission facility-related costs, which are more likely related to average loads than to peak loads; and generation-related costs, which arise because of transmission losses and transmission constraints and which are related to the locations of generators and loads.
- *Generation services*, which are the provision of electrical energy, regulation, operating reserves, reactive power, and blackstart. The costs of these services are related to customers' loads and locations.

These states have established unbundled rates designed to recover separately the costs allocated to each service (*e.g.*, separate customer, demand and/or energy charges for customer, distribution, transmission, and generation services). Unbundling of generation costs in particular is required if consumers have the option of purchasing energy from a source other than their distribution utility. However, even utilities in non-restructured states have seen the business value in understanding their unbundled costs, even if their rates remain bundled.

4.2. The Marginal Costs of Power Supply

Because the costs of non-competitive services (customer services, distribution, transmission, reactive power, and blackstart) are either fixed, set by regulation, or both, the marginal costs

that are of most interest from the standpoint of time and location variation are those of the competitive generation services (energy, regulation, and reserves). These latter marginal costs include those that arise from transmission losses and constraints.

Because electricity is essentially non-storable, it must be generated at the time that it is demanded. Because consumers' demands for electricity vary considerably over time, on an hourly, daily, and seasonal basis, the system operator must continually vary the output of generators so that their output matches load.

Generation technologies differ substantially in terms of their capital and operating costs. These technologies include capital-intensive *baseload* plants that are designed to run continuously at low operating costs, and *peaking* plants that are relatively inexpensive to build but that have high operating costs during the relatively few hours of the year that they run.

To minimize costs, the load at any moment in time is served first by the generating units with the lowest incremental energy costs, and last by the generating units that have the highest incremental energy costs.[11] The "incremental energy cost" of a generating unit is defined as the change in fuel costs and other variable costs that accompanies a small increase (e.g., 1 MW) in its output. In general, hydroelectric, nuclear, and coal units have relatively low incremental energy costs, while gas- and oil-fired units have relatively high incremental energy costs.

The power system's "marginal energy cost" is defined as the change in the power system's fuel costs and other variable costs that accompanies a small increase (e.g., 1 MW) in system load. Marginal energy cost will equal the incremental cost of the power system's marginal generator, where the marginal generator is usually the generator that has the highest incremental energy cost of all of the generators that are operating above their minimum output levels. Marginal energy cost is important because it is a major determinant of the economic value of electricity, of the efficient market price of electrical energy, and of the actual market prices of electrical energy.

In theory, the efficient price of electrical energy at any point in time should equal the power system's marginal energy cost plus a premium that reflects scarcity (if any) of generation supply relative to load. This scarcity premium should depend upon the value of reliable electric service to customers.[12] As a practical matter, because loads and generator availability change over time, different combinations of generators will serve load at different times; so marginal energy costs, efficient energy prices, and actual market prices will change over time, sometimes substantially from one hour to the next.

Marginal energy costs can also vary substantially by location. This occurs because electricity must be transported from generators to consumers over a network of power lines. In the course of transport, some power is lost as waste heat; so marginal energy costs vary among locations according to the marginal transmission losses among those locations. Furthermore, all transmission lines have capacity limits. When these limits are reached, there is a separation of

---

[11] For simplicity, the text ignores the effects of various operating constraints (like ramping constraints) upon merit-order dispatch. Nonetheless, the text describes a reasonable approximation to reality.

[12] The premium should approximate: a) the change in expected unserved energy that accompanies a small load change; times b) the excess of the marginal value of electricity to consumers over the power system's marginal energy cost. If operating reserve markets are efficient, the prices of operating reserves can serve as a basis for determining this premium.

the markets on opposite sides of the constrained transmission facilities, so that marginal energy costs and efficient prices may be quite different on the opposite sides.

The marginal costs of power supply arguably also include costs that are not borne directly by electricity market participants. These can include the environmental costs associated with fuel extraction (e.g., damage to wildlife), fuel combustion (e.g., air pollution), fuel disposal (e.g., nuclear waste), plant siting, and transmission line siting.

4.3. Marginal Costs and the California Market Design

The California market operates more or less as it was designed in the mid 1990s; except that, with the collapse of the Power Exchange in the power crisis year 2001, California lacks a day-ahead market. Some notable features of the California market are:

- It has a same-day hourly market, but no day-ahead market. The lack of a day-ahead market reduces market participants' hedging options and, more seriously, makes system operation more difficult.
- It has sequential auctions by which the energy market clears first, and then each of the ancillary service markets clear in succession. This leads to inconsistent prices among the various generation service markets.
- It has zonal pricing of electricity. This means that transmission congestion between zones is managed through a price mechanism, while transmission congestion within zones is managed through non-price mechanisms.

Since 2002, the California Independent System Operator has formally proposed several major reforms of the market. The current proposal is called the "Market Redesign and Technology Upgrade" and has the following important features:

- It adds a day-ahead market.
- It has simultaneous auctions of energy and ancillary services.
- It has nodal pricing of electricity, so that transmission congestion will be generally managed through a price mechanism that makes no distinction between inter-zonal and intra-zonal congestion.

From the perspective of efficient retail rate design, the reform has two major implications. First, the day-ahead market will provide an objective basis for: a) setting prices for day-ahead retail pricing programs; and b) forecasting and calling service curtailments that depend upon day-ahead information.[13] Second, the energy prices that will be relevant for each customer will be the nodal price applicable to that customer's location; so different customers can in principle face different prices and/or different curtailment calls. These locational differences in the treatment of customers can be particularly important when there are reliability problems in

[13] Southern California Edison apparently harbors significant doubts about the continuing usefulness of the price signals of the California ISO markets. In forecasting marginal energy costs, it uses a mixture of price forecasts and marginal cost forecasts, where the explicit purpose of using the marginal cost forecast is "to account for the declining liquidity of the market…" While basing 100% of its forecast for 2006 on market prices, it bases only 47% of its forecast for 2008 on market prices. See Southern California Edison [2005, pp. 23-24].

certain load pockets (*e.g.*, San Francisco) that can be relieved only by curtailments in those load pockets, and cannot be relieved by curtailments elsewhere.

5.0 Developing Efficient Retail Rates

To provide appropriate incentives for demand response, retail prices need to reflect the variations in the marginal costs of electricity across time and locations. Retail rates in which prices vary by time period have the price variations across time; but the types of time-varying rates differ from one another along several dimensions. These dimensions include:

- *The fineness of time increments*. Prices may vary hourly, or they may vary by on- and off-peak periods, or they may vary seasonally.
- *Price notification period*. Prices may be announced a year in advance of taking effect, a season in advance, a day in advance, or an hour in advance.
- *Price updates*. A program may, under specified conditions, allow the load-serving entity (LSE) to change prices on shorter notice than indicated by the price notification period. For example, a time-of-use program with prices that are announced a year in advance might allow the LSE to update prices for up to twenty hours with only day-ahead notice.
- *Limits on price variability*. A program may include a guarantee that prices will not exceed a specified ceiling.

This section describes a set of pricing principles for designing retail rates that provide appropriate incentives for demand response. It is divided into four parts. The first part describes the various types of time-varying rates. The second part presents pricing principles, including mathematics for calculating prices. The third part discusses customer choice issues. Finally, the fourth part lists information requirements and potential information sources.

5.1. Types of Time-varying Retail Rates

Conventional utility prices, particularly for small customers, are "flat rates" that are constant in all hours of the year. Time-varying rates, by contrast, offer a spectrum of price accuracy and price certainty. In essence, the more accurately prices reflect actual power system conditions, the more uncertain they will be; while the more certain prices are, the less accurately they will reflect actual power system conditions. Accurate prices provide appropriate signals of the value of demand response, but (because of uncertainty and variability) they are generally unattractive to consumers.

The two major categories of time-variation are as follows:

- Real-time pricing (RTP)
- Time-of-use (TOU) pricing (including time-of-day and seasonal pricing)

RTP prices reflect nearly current power system conditions, so they have the greater price accuracy and the lesser certainty. TOU prices, by contrast, are set far in advance of the periods to which they apply; so they have less accuracy and greater certainty. For the TOU options (and for flat rates), a dynamic pricing feature may be added that, on short notice, sets prices according to power system conditions, while most hours' prices are set in advance. This pricing option is called critical-peak pricing (CPP).

5.1.1. Real-time pricing

Real-time pricing programs have prices that change hourly.[14] At the extreme, customers could pay actual real-time market-clearing prices; but because these prices are not known until after the fact, few consumers are likely to find such a rate attractive due to the risks of price variability and uncertainty, and such a scheme would provide poor incentives for efficient demand response.[15] In RTP programs at regulated utilities in the U.S., prices tend to be set a day in advance for one market clearing, and then set an hour in advance for a second market clearing.

For those customers who are willing to face RTP prices but want limits on the uncertainties that they face, there are a number of mechanisms for limiting uncertainty. They include:

- *Price caps* that set a limit on how high prices may rise. Customers would pay a premium or option price to cover suppliers' costs of guaranteeing the cap.
- *Price collars* that put both floors and ceilings on prices. By paying extra for power during low-price periods, customers would implicitly pay a premium or option price to cover suppliers' costs of guaranteeing the cap.
- *Price hedges* by which customers contract to buy a portion of their expected load through a forward contract for a block of power during some future period at a fixed price. For example, a customer with a typical usage level of 5,000 kW may enter a contract to buy a block of 4,000 kW in every daytime hour on weekdays during a particular month in the future, at a specific forward contract price. With a guaranteed price for 80% of the expected load, the customer would face price risk only on the remaining 20%. This arrangement is analogous to a two-part RTP tariff under traditional regulation, as described below. Such forward contracts for price hedging may be obtained from the customer's energy supplier, or from a financial agency or commodity market.

Price caps and price collars blunt price signals: in the high-price hours when load reduction is needed most, the customer lacks the full incentive to reduce load. Price hedges, by contrast, limit the financial risk repercussions of high prices but still give customers the full economic incentive to respond to power system conditions. For example, a 1,000 kWh load reduction in an hour in which the price is $0.50 per kWh will reduce the customer's bill by $500, regardless of the price hedges that the customer may have.

Note that an unbundled version of RTP with price hedging also resolves certain issues that have concerned California utilities in the past. That is, since unbundled transmission and distribution (T&D) services are priced separately based on current consumption, if RTP consumers expand beyond their contractual CBL, they will pay T&D charges on that incremental usage.

A separate issue exists regarding the recovery of other non-bypassable charges, such as above-market generation costs associated with power contracts signed during the 2001 electricity crisis. As discussed in Section 5.2.1.2, two principle cost recovery alternatives are available.

---

[14] In the U.S., RTP is hourly. In some other countries, RTP is half-hourly.

[15] We are aware of RTP-like pricing products offered by competitive retail providers in Texas, where no day-ahead market exists. Consumers that accept these products appear willing to work with short-term price-tracking information and price notices made available to them by their energy provider.

One is to recover those costs through a rate that is applied to all current usage. Another is to recover the costs through a rate that applies only to a baseline level of usage, so that incremental or decremental usage (relative to the baseline) is priced according to marginal costs or market prices. The disadvantage of the first approach is that applying the rate to all current usage generally distorts prices away from actual marginal costs – customers would pay too much for usage increases and would receive too large a bill reduction for usage reductions. By contrast, the second approach bases each consumer's contribution to the non-bypassable costs on an historical usage level that is arguably closer to the level in place at the time that those costs were incurred; and changes in usage are priced close to market values.

5.1.2. Time-of-use pricing

Time-of-use (TOU) prices are fixed within each pricing period. Pricing periods may be defined according to time of day, day of week, or season. Utilities may establish demand prices as well as energy prices that vary by time of day. Specific types of TOU rates include the following:

- *Time-of-day rates*. Under these rates, prices vary by time of day and day of week, but not by season. There might, for example, be a peak period from 6 am to 10 pm every weekday, and an off-peak period in all other hours.
- *Seasonal rates*. Under these rates, prices vary by season but are constant within each season. For example, there may be a summer rate in June through September, and a non-summer rate in all other months.
- *Seasonal time-of-day rates*. Under these rates, prices vary by time of day, day of week, and season. There could be one price in the summer peak hours, another price in the summer off-peak hours, a third price in the non-summer peak hours, and a fourth price in the non-summer off-peak hours.
- *Day-type time-of-day rates*. With day-type TOU, all prices are announced well in advance, but they apply to days according to information available at or close to real time. For example, a day-type TOU program might announce three sets of peak and off-peak prices—for a cool summer day, a hot summer day, and a sizzling summer day—each of which are defined by temperatures forecast a day ahead. Customers then know well in advance what prices that they will face; they might even be guaranteed that there will be no more than (say) ten sizzling summer days; and the prices that they actually face will reflect current conditions.

TOU and seasonal TOU rates are common throughout the industry. Because TOU rates are always announced months in advance, they induce customers to respond to forecast power system conditions, not to actual power system conditions. That response is better than nothing: it can provide some load relief, for example, in the summer on-peak period.[16]

[16] In a project for the California Energy Commission to evaluate the real-time energy metering project in which California utilities installed advanced metering equipment for all customer accounts of size greater than 200 kW, we estimated those customers' price responsiveness to PG&E's and SCE's TOU demand and energy rates. We found that approximately 10% to 20% of customers in nearly all SIC groups showed significant signs of load reductions in summer peak periods (during which substantially higher

It would be even better, however, if TOU rates could be set closer to the times to which they apply, based upon more current information on power system conditions. Day-type time-of-day rates do have such an information update, and thus have properties similar to those of the CPP rate described next.

5.1.3. Critical peak pricing

Critical peak pricing (CPP) is a pricing option that is combined with either a TOU or flat rate. With CPP, customers face prices that, for most hours, are announced well in advance; but in some critical hours, when load relief is most valuable, the customer will face high prices that reflect expected power system conditions. Critical hours may be defined by load levels, operating reserve levels, price levels, or ambient temperature.

CPP programs typically limit customers' price risk. Specifically, CPP programs may be designed so that there are a maximum number of CPP events or hours per year (*e.g.*, fifteen events, or 80 hours per year), with pre-specified CPP prices that are announced in advance. CPP prices may in principle be designed to vary hourly with power system conditions, or have more than one fixed level, such as "critical" and "super-critical." Such enhancements add accuracy at the expense of simplicity.

Customers may be attracted to CPP because it appears less complex than full-time RTP, since prices do not vary during most hours of the year. Thus, customers only occasionally have to worry about determining what their peak-period prices will be, and about deciding whether and how to respond to them. Furthermore, because CPP programs reduce energy suppliers' price risk and expected cost to serve, customers will generally pay lower prices in non-CPP hours than would be the case for a fully guaranteed-price product.

5.2. Pricing Principles

The following discussion of pricing principles is broken into two parts. The first part discusses major considerations in setting retail prices for unbundled competitive generation services: real energy, regulation, and operating reserves.[17] To the extent that transmission losses and constraints affect the locational wholesale prices of the competitive generation services, our discussion of retail prices refers to the wholesale prices applicable to the retail customer's location. The second part discusses the services (e.g., reactive power, black start, transmission, distribution, and customer services) with prices that must be set through non-competitive processes.

All of this discussion applies equally to customers of all rate classes. For purposes of designing efficient retail rates, there are (in principle) only three significant differences among customers. First, customers differ in their load patterns. Second, customers differ in their responsiveness to

---

TOU demand charges applied) relative to the same time periods in non-summer months. See Christensen Associates Energy Consulting [2005].

[17] In the absence of resource adequacy requirements, any market for capacity or for planning reserves is a derivative of the markets for real energy, regulation, and operating reserve services. If there is a resource adequacy requirement, capacity or planning reserves will have values that also depend upon the manner in which the resource adequacy requirement is imposed.

price signals. Third, customers differ in the sizes of their loads relative to the program costs. As indicated by equation 5 below, the first consideration means that different customer groups will have different efficient prices. The second consideration should surely influence the marketing of variable-price programs, and may also imply that different methods of cost recovery are appropriate for different customer groups. The third consideration also influences program marketing. But for all customers, the principles described below are identical.

5.2.1. Retail pricing of unbundled competitive generation service

The key to the success of a variable-price program is in setting the appropriate prices for generation services, including fuel and capacity costs. The appropriate retail price structure will depend in part on the nature of the wholesale market design, particularly the services (e.g., energy, capacity) that are included in the market and the ways that those services are priced. For example, if energy, regulation, and reserves are subject to market-based pricing, it may be appropriate to recover generation costs only through retail energy charges. If there is also a capacity market, it may be appropriate to recover capacity costs through retail demand charges.

In the absence of capacity markets, our view is that energy prices (*i.e.*, based on kWh of energy consumption), rather than demand charges (*i.e.*, based on kW of maximum demand), should be used to recover *all* costs of generation services. If the energy prices are set appropriately, they will ultimately reflect both the variable operating costs and the capacity costs that the customer imposes on the system, while maintaining a clear and appropriate price signal at the margin. If California institutes a capacity market that has prices analogous to demand charges, then it may be appropriate to recover those costs through retail demand charges that mirror the capacity charges, although those costs could still be recovered through appropriately designed energy prices that reflect the probability of incurring capacity charges. That is, the component of energy prices that recovers the capacity charges would not be constant across all hours or time periods, but would be higher during time periods that are more likely to create a need for additional capacity.

Demand charges have the chief advantage of stabilizing the recovery of fixed capacity costs. However, they tend to distort customer incentives. Consider, for example, a rate that has a relatively high demand charge to recover capacity costs and a relatively low energy price to recover fuel costs. In this case, a customer who experiences one hour of abnormally high usage faces a very low incremental energy cost for the remainder of the billing period. In contrast, a customer who has a high load factor may face a very high cost of increasing its usage above its typical demand level, even though the market price of energy may be very low.

In addition to avoiding the use of demand charges for generation services, some forms of energy rates are less effective in communicating accurate price signals. Included in this list are block rates (inclining and declining) and hours-of-use rates. In both cases, these rates provide incentives that are related to desired outcomes that do not necessarily include reflecting marginal power costs. That is, inclining block rates provide customers with an incentive to reduce total usage; declining block rates provide customers with an incentive to increase total usage (because capacity is abundant) or to address equity concerns for space heat customers; and hours-of-use rates provide an incentive to improve load factor. However, these rates do not provide price signals that reflect current market conditions. For example, an inclining block rate may provide a relatively high energy price to customers at higher usage levels at times

when the market price of energy is quite low. This leads to an inefficient reduction in the amount of energy consumed (i.e., the customer was willing to pay more for the incremental energy than it cost to produce, yet the energy goes unconsumed).

Block and hours-of-use rates can also be difficult for customers to understand. For a block rate, the customer must keep track of their current usage level and their expected usage for the remainder of the billing period to correctly assess the price that they will pay for changing their usage level. The problem is more complex with an hours-of-use rate, in which the customer must keep track of both their usage level and their maximum demand in order to be able to determine the price that they are paying.

5.2.1.1. Efficient pricing

This section provides a mathematical description of the appropriate method for calculating any of the retail energy rates described in Section 5.1. In fact, the key to offering consumers a choice among rate options is to use the same fundamental algorithm for each rate, accounting for differences in price risk implied by the features of the rate.

The efficient price for retail energy usage depends upon the time period T (*e.g.*, summer on-peak) and the customer group j (*e.g.*, medium-sized commercial and industrial customers) to which the price applies. This efficient price can be expressed as follows:

$$P_{Tj}^{*} = \frac{\sum_{h \in T} E\left\{Q_{hj} * \left[P_h^{EN} + RR_h^{REG} * P_h^{REG} + RR_h^{SR} * P_h^{SR} + RR_h^{NSR} * P_h^{NSR}\right]\right\}}{\sum_{h \in T} E\left\{Q_{hj}\right\}} \quad (5)$$

where the variables are as follows:

P_{Tj}^{*}	the efficient price for customer group j in time period T ($/MWh)
E	the expectation operator
Q_{hj}	the load of customer group j in hour h (MWh)
P_h^{EN}	the wholesale market price of energy in hour h[18] ($/MWh)
P_h^{REG}	the wholesale market price of regulating reserves in hour h ($/MW/hr)
P_h^{SR}	the wholesale market price of spinning reserves in hour h ($/MW/hr)
P_h^{NSR}	the wholesale market price of non-spinning reserves in hour h ($/MW/hr)

[18] For simplicity, we assume that all relevant marginal costs are reflected in market prices. These include environmental costs and scarcity values (i.e., outage or capacity costs). Tradable emissions permits provide a market mechanism to value the environmental impacts of generation, the cost of which the generators would recover in the energy market. As noted by Oren [2005], inauguration of a capacity market would reduce market energy prices and call for somewhat different treatment than indicated in the text.

RR_h^{REG} the regulating reserve requirements ratio in hour h (unitless)

RR_h^{SR} the spinning reserve requirements ratio in hour h (unitless)

RR_h^{NSR} the non-spinning reserve requirements ratio in hour h (unitless).

The reserve requirements ratios are ratios of reserve requirements to load. In California, spinning reserve requirements and non-spinning reserve requirements are each typically about 3% of load; so these reserve requirement ratios would each typically be about 3%.

Equation 5 says that efficient retail energy prices depend upon expectations of hourly loads and wholesale prices. For each pricing period, the efficient price is a load-weighted average of the hourly wholesale competitive prices for that time period, where the weights are the hourly loads. For regulation and reserve services, prices are further weighted by the ratios of reserve requirements to loads. Because these ratios are small, this gives regulation and reserve services little weight relative to energy prices.

Equation 5 applies to fixed-price and variable-price products. For a fixed-price product (*e.g.*, seasonal pricing), the time period T will be large, perhaps a year or more. For a variable-price product, the period will be shorter: for TOU, it will encompass TOU periods of dozens of hours in each week and many weeks in each season or year; for RTP, it will encompass a single hour; and for CPP, it will encompass critical hours.

Although it may appear that the retail price for a given time period would simply be a weighted sum of the expected wholesale prices for the relevant hours, that interpretation is incorrect because of the correlation between loads and wholesale prices. That is, both loads and wholesale prices are likely to be higher than expected on hot days than on cool days. Such a positive correlation between hourly loads and wholesale prices means that equation 5 implies that the efficient retail price will be slightly higher than the simple load-weighted sum of the expected wholesale prices. The efficient retail price will be particularly higher than the expected wholesale price for those customer types whose loads tend to be strongly correlated with wholesale prices, such as commercial customers whose loads tend to rise on hot days, than for those customer types whose loads are relatively uncorrelated with wholesale price, such as large industrial customers.

In the case of a time-of-use rate, equation 5 can be applied separately for each pricing period. For example, one would forecast the expected product of loads and prices and the expected loads during the summer peak hours, and then derive the summer peak price by inserting the values in equation 5. The same procedure would be followed for the other pricing periods.

Equation 5 can also be used to set the critical price and non-critical prices within a CPP rate. This requires a forecast of the hours that are expected to be called. Equation 5 would then be applied separately to the critical hours and the non-critical hours. Equation 5 provides an economic basis for the fact that non-critical energy prices in a CPP rate are less than energy prices in a standard rate. The price difference is due to the fact that in the CPP rate, high-price (*i.e.*, critical) hours are excluded from the non-critical energy price calculations.

5.2.1.2. Cost recovery issues[19]

Prices equal to marginal costs—or equal to weighted averages of marginal costs, as in equation 5—will sometimes collect less revenue than is required to recover costs and will sometimes collect more revenue than is required to recover costs. Under-collections will generally occur in the "bust" part of a business cycle, when there is surplus generation capacity, or when price inflation is unexpectedly low; while over-collections will generally occur in the "boom" part of a business cycle, when there are generation capacity shortages, or when inflation is unexpectedly high.[20] Competitive firms will lose money during periods of under-collection and will make high profits during periods of over-collection. By contrast, regulated firms subject to cost recovery constraints must somehow have their prices or price structures adjusted so that no significant under-collections or over-collections occur.[21]

A number of pricing strategies have been proposed and used to achieve the most efficient prices while achieving adequate revenue recovery. In this context, "most efficient prices" are those that cause consumption to be as close as possible to what it would be if prices equaled marginal costs. The fundamental conclusions in the technical literature are that efficient pricing strategies involve a combination of Ramsey pricing and non-linear pricing. "Ramsey pricing" refers, essentially, to a procedure by which prices diverge from marginal cost according to the inverse of customers' price responsiveness.

"Non-linear" pricing refers to pricing structures that have multiple prices in each time period. This is the opposite of uniform pricing, whereby a single price per kWh applies to all consumption during a period. Examples of non-uniform, or non-linear, prices include the following:

- *Block tariffs* have different prices for different levels of consumption. Declining block tariffs have a high price on low consumption and a lower price on higher consumption: these address under-collections. Inclining block tariffs have a low price on low consumption and a higher price on higher consumption: these address over-collections.
- *Two-part tariffs* have an energy charge and an access charge. The energy charge can have prices close to marginal cost. The access charge is a fixed charge, related to each customer's historical consumption, which makes up the difference between marginal cost revenues and cost recovery requirements on the customer's historical consumption.
- *Multi-part tariffs* can mix various combinations of energy, demand, and access charges, and can include block tariffs.
- *Self-selecting menus of two-part tariffs* allow consumers to select from various combinations of access charges and energy prices. An example of such self-selecting

---

[19] Brown and Sibley [1986] and Wilson [1993] provide comprehensive technical discussions of optimal pricing, with special application to public utilities that have an allowed revenue requirement.

[20] At the present time, California has an under-collection problem due (in part) to "legacy costs," notably including the costs of the high-priced power contracts signed by the State of California at the height of the power crisis of 2000-2001.

[21] Balancing account ratemaking is used to assure that under- or over-collections in one period are resolved through rate changes in a subsequent period. The text's discussion of prices and price structures applies to pricing in both periods, given whatever the cost recovery targets are for those periods.

menus is provided by cellular telephone plans that offer various combinations of monthly charges and allowable peak-period minutes.

For example, the largest RTP programs currently operating under traditional regulation utilize a *two-part* design.[22] This feature allows RTP prices to closely track changes in wholesale prices, while at the same time reducing risk to both the utility and the customer. It allows the utility to recover the amount of revenue allowed by the standard tariff that the customer would otherwise face, and limits bill changes if RTP prices reach unexpected levels. In summary, unbundled two-part RTP: gives customers an incentive to reduce load at times of high RTP prices; insures customers against large bill increases even if they do not respond; assures suppliers of full cost recovery on baseline usage levels; and gives customers an opportunity to buy additional power during low-price hours and to effectively sell power back to the utility during high-price hours.[23]

As another example, multi-part tariffs have the fundamental objective of applying to consumers' incremental usage decision a marginal price that has the smallest possible mark up over marginal cost, with revenue recovery in excess of marginal costs occurring on non-incremental portions of consumers' usage. This is an application of Ramsey pricing, as the mark up on infra-marginal usage, which is relatively non-price responsive, exceeds that on the more price-responsive incremental usage.

Cost recovery constraints pose some challenges for getting prices and incentives right for demand response. Nonetheless, the foregoing approaches can be used to minimize the distortions that are inevitably created by these constraints. In short, cost recovery constraints are a challenge, but not a major challenge, in providing efficient incentives for demand response.

5.2.2. Retail pricing of non-competitive services

Customer service costs should generally be recovered through customer charges. This is because these costs are generally related to the number of customers within each class.

Unbundled distribution costs should be recovered through a combination of customer and demand charges, provided that demand metering is available. This will help ensure the recovery of fixed distribution costs, which is the primary concern for this category of costs. The use of demand charges (as opposed to energy charges) to recover these costs both stabilizes the revenue stream and more reasonably matches the costs that customers impose on the network. For example, the distribution system must be designed to the maximum non-coincident peak demand of each distribution area, so charging customers based on a ratcheted non-coincident demand reflects the cost that they arguably impose on the network.

Transmission, reactive power, and blackstart costs should be recovered through the least distorting charges. In restructured markets, recovery has tended to be through energy charges.

---

[22] These include, for example, the RTP programs offered by Georgia Power Company and Duke Power Company.

[23] Appendix B provides a more comprehensive description of customers' bills and demand-response incentives under an unbundled and financially hedged RTP product.

5.3. Customer Choice Issues

This section looks at the extent to which customers should be allowed to choose among pricing options and at customer eligibility for variable pricing options.

5.3.1. The extent of customer choice

Customers have different preferences for price certainty. Some customers—particularly those who are most able to respond to price changes—are willing to face prices that can change with little notice and that may vary substantially from hour to hour and day to day. Other customers—particularly those who are unable to shift loads among time periods—prefer prices that are stable and announced well in advance; and they would be willing to pay a premium to avoid price variations and price uncertainty.

It would be an acceptable outcome if all customers, given the choice, were willing to pay the full costs associated with having fixed retail prices while maintaining system reliability. In this extreme case of no price-responsive load, the role of demand response would be replaced by maintaining a large amount of reserve capacity in the system, for which customers have expressed a willingness to pay. Under traditional ratemaking, consumers are not given the choice to express their willingness to pay for the additional capacity needed to meet non-responsive demand reliably. As a practical matter, when market conditions are sufficiently tight, the premium associated with fixed-price products will encourage some customers to accept a variable-price product if such a product is available. Customers' acceptance of variable-price products will produce a balance between the need for price-responsive load and generation capacity.

While customers should be allowed to choose from a range of products, some research has indicated that customers are less likely to make a rational choice (or be satisfied with their decision) when the number of alternatives is too large.[24] To reduce the potential for customer choice "paralysis," we propose that the alternatives be limited to a small number of programs that are clearly differentiated along the risk/price level continuum. For example, a portfolio of products for commercial and industrial customers might include a *seasonal time-of-use* rate; a *CPP* rate that provides fixed prices in most hours of the year, but some opportunity to take advantage of demand response; and a *real-time pricing* rate that allows the customer to minimize the risk premium that they pay for price certainty while maximizing their opportunity to take advantage of varying market prices. Of course, there can be a wide range of options in a competitive market, as competitors are free to develop their own product portfolios.

The choice of the default rate is important because of status quo bias. "Status quo bias" refers to the fact that consumers tend to remain on whatever rate that they happen to have, rather than switch to another rate, even if the other rate will save them money or provide them with other benefits.[25] This is a transition issue that can be addressed in Phase II. It is also worth noting

---

[24] See Schwartz [2004].

[25] The term "status quo bias" appears to originate with Samuelson and Zeckhauser [1988], who applied economics, psychology, and decision analysis research to explain consumer decision-making relative to an existing alternative (status quo). They and others have found an apparent "endowment effect" in which the value of a product or of a decision (*e.g.*, subscription to an existing electricity tariff), increases

that studies of customer price responsiveness and participation in dynamic pricing and demand response programs suggest that a relatively small percentage of customers (*e.g.*, 10 to 15%) tend to be most price-responsive and thus most likely to choose a time-varying rate option.

5.3.2. Customer eligibility for variable pricing options

Customer eligibility for each variable pricing option will depend upon the available metering. In general, expenditures on advanced meters (capable of recording hourly usage) tend to be economic primarily for those customers with the largest and most price-responsive loads.[26] In California, most customer accounts with maximum demand in excess of 200 kW have advanced meters installed that are compatible with dynamic pricing products such as RTP, CPP, and day-type TOU. In addition, the business case for expanding advanced metering to all consumers is under active consideration by the utilities and regulatory agencies.

Although many customers will not currently be eligible for variable pricing options, even non-participating customers can benefit indirectly from the availability of these options to other customers. The indirect benefits arise from the fact that, if properly priced, the variable pricing options will provide cost-reduction benefits and reliability benefits to both participating and non-participating customers.

5.4. Information Requirements

This section lists the information required to design variable-price products and to set their prices.

5.4.1. Forecasting customer loads

The hourly customer loads Q_{hj} may be obtained from historical metered data for individual customers, samples of customers, or entire customer classes for which interval metered data are available. Historical relationships between hourly loads and wholesale prices may also be examined to develop estimates of the risk component of efficient retail prices. Note that the load values Q_{hj} should be *after* customer response to price, not before response. For re-pricing ongoing programs, historical data may be used. For pricing a new rate, a demand model combined with assumptions on price responsiveness may be used to simulate load response. Note that pertinent information may eventually become available as a result of the CPUC-mandated protocols for estimating load response.[27]

---

once it has been acquired or made. Thereafter, consumers demand more to give it up than they would be willing to pay to acquire it.

[26] The California Energy Commission, the California Public Utilities Commission, and Governor Schwarzenegger have all strongly supported making available to *all* customers the "advanced metering and communications systems capable of supporting time-based rates…" See California Energy Commission [2005, pp. 72-73].

[27] See California Public Utilities Commission [2005, p. 12].

5.4.2. Forecasting energy and reserve prices

The prices P_h^{EN}, P_h^{REG}, P_h^{SR}, and P_h^{NSR} can be forecast by methods such as: a) forecasting supply and demand curves and thereupon estimating market-clearing prices; and b) deducing likely prices from past price histories.

5.4.3. Determining reserve requirements ratios

Information on reserve requirements may be obtained from the California ISO.

5.4.4. Estimating demand response to price

An indicator of the natural limits on the expected amount of responsive demand is provided by consumers' cost of curtailing load. That is, demand response does not come without cost. Substantial empirical evidence from analyses of CPP, RTP, and DR programs indicates that some consumers are willing to curtail usage somewhat, on occasion, if they are compensated sufficiently.[28] However, many consumers find it difficult or costly to modify their usage at short notice. These findings suggest that the aggregate demand curve is relatively steep, and the aggregate price elasticity of demand is low; so a large price increase is required to produce a modest amount of demand response. It is possible that the usage flexibility of some customers might be improved through the application of enabling technologies, such as pre-programmed energy management systems tied to daily price communication equipment; however, the cost of those systems also adds to the cost of load curtailments.

5.5. Implementation Issues

This section briefly discusses metering and self-selection bias issues

5.5.1. Metering

In order for customers to be able to benefit from variable-price rates, metering must be in place that is capable of recording their usage changes during variable-price time periods (*e.g.*, critical-peak periods). For most large customers (*e.g.*, those with maximum demand greater than 200 kW), the required metering technology is already in place. In contrast, the issue of the cost effectiveness of advanced metering for small customers has been the subject of recent utility and regulatory evaluation, where the potential benefits associated with demand response are being combined with other cost saving benefits to determine overall cost effectiveness.

5.5.2. Adverse selection

The efficient prices of equation 5 can be calculated for groups of customers or for individual customers (i.e., "groups" of one customer each). For rates that cover multiple time periods, like

[28] See Braithwait and O'Sheasy [2001] regarding RTP price response, and Neenan *et. al* [2003] for discussion of load response in a DR program. Braithwait [2000] and Charles River Associates [2005] report estimates of customer price responsiveness to CPP rates.

TOU rates, setting prices according to a class-wide average load profile will lead to self-selection bias. For example, those customers with relatively high usage in off-peak periods can reduce their bills merely by shifting from a flat rate to a TOU rate, without altering their usage profiles; while customers with relatively high usage in on-peak periods will remain on the flat rate if they can. This means that customers will choose the rate options that are least profitable (or most money-losing) for the LSE. This can lead to revenue attrition for the LSE.

In anticipating potential issues in implementing a menu of default and optional rates, it is useful to consider an "ideal" case in which consumers receive individually customized prices based on the cost of serving their actual usage pattern. If the customer load data were available, equation 5 could be used to set customer-specific prices that properly account for the cost to serve each customer.[29] Customer-specific pricing has the following advantages:

- It is "fair," in that the same pricing method is applied to each customer, and the resulting prices reflect the expected cost to serve the customer; and
- Revenue attrition from customer self-selection of optional class-wide rates is minimized.

The self-selection (or adverse selection) problem is eliminated when each customer is priced based on its own load profile. If a customer with more than average off-peak usage faced a choice between a relatively low flat guaranteed price and the CPP rate, individually customized prices would leave the LSE indifferent to the customer's choice of rate. (Recall that the pricing method properly accounts for both the risk associated with the flat rate and the expected benefits of the load response that the CPP rate enables.) Individually customized prices have an additional attractive property to the extent they are updated over time based on consumers' actual consumption patterns. In that case, even consumers facing fixed, non-dynamic rates have an indirect incentive to respond to publicized conditions of high wholesale power costs. That is, if they reduce current usage during such periods, their fixed price in future years, as calculated through the methods of equation 5, will be lower than if they did not respond.[30]

Because of metering technology, tradition, and cost and data limitations, the ideal of customer-specific pricing is unlikely to be achieved for all customer classes in the near future. Thus, the issue of self-selection with optional class-wide rates must be addressed. The potential for customer self-selection to produce revenue attrition increases as more diverse customers are grouped together, and is reduced when the default tariff is already time-differentiated, as with the existing TOU demand and energy rates for large customers in California. When rate alternatives are developed for classes of customers, a risk premium must be added to equation 5 to account for the risks associated with self-selection.

---

[29] Customer-specific pricing may appear impractical at first glance. However, two-part RTP programs involve the development of customer-specific baseline loads for each participating customer. In addition, utilities that have offered fixed-bill products calculate customer-specific offers based on each customer's historical monthly usage patterns. Finally, competitive service providers must tailor their price offers to either individual customers or relatively narrow customer types to minimize the risk associated with quantity uncertainty.

[30] See Glyer [2000].

6.0 Developing Efficient Curtailable Service Programs

Due to the unique features of electric power systems, maintaining reliability can sometimes require actions on a time frame too short for demand to respond to price signals. Thus, even with efficient pricing mechanisms in place, curtailable service programs can provide a valuable form of demand response under some power system conditions. As described in this section, the design and pricing of efficient curtailable service programs follows the same basic principles as those for efficient rate design.

A curtailable service program gives the power system operator the right to curtail a customer's load under certain pre-specified conditions. The conditions may depend upon events within the power system, like real-time reserve margins or emergency actions that the system operator would need to take without the curtailments. The conditions may also depend upon factors like ambient temperature.

A curtailable service program can also give an LSE the right to curtail load under certain conditions. In particular, LSEs can gain the right to curtail customer loads in response to high wholesale prices, which would enable LSEs to either cut their costs or sell power into the wholesale market.

A curtailable service program places limits on the curtailments that can be called by system operators and LSEs. These limits include minimum advance notice periods, length of curtailments, number of curtailments, and so on.

6.1. General Principles for Curtailable Service Program Design

There are two basic options for designing curtailable service programs.

In *traditional* curtailment programs, the customer receives a pre-determined credit for participation in the program. This participation credit is usually in the form of reduced demand or energy charges. Almost all present curtailable rate programs are of this type. Some programs' credits depend upon customers' actual performance during curtailments. Few programs' credits depend upon the frequency with which curtailments are actually called.

In a *performance-based* program, the customer receives a (smaller) pre-determined credit for program participation and separate payments for actual curtailments. As with traditional programs, the participation credit is usually a discount in demand or energy charges. The payment for actual curtailment is made sometime after each curtailment, based on energy prices announced prior to the curtailment event.

6.1.1. The benefits of curtailable load

Curtailable load provides a power system with an *insurance value* and an *operating value*.

The insurance value arises from the operating reserves that are provided by curtailable load. Curtailable loads are like other operating reserves in that they can help resolve power imbalances and thereby improve system reliability even if load is not actually curtailed.

The operating value arises from the improved power system reliability and/or generation cost savings that occur when load is actually curtailed. The reliability benefit is manifest if the curtailment helps reduce the chance that the power system will incur costly reliability problems

like rotating blackouts or system collapse. The cost savings occur if the curtailment allows the power system to avoid operating costs that are higher than the retail power prices paid by curtailed customers.

From the perspective of power systems, curtailment programs' values are highest when: the right to curtail is firm; the notice period is short; the MWs curtailed can be high; curtailments can be numerous; curtailments can be frequent; and the duration of curtailments can be long.

6.1.2. The costs to customers of participation in a curtailment program

Customers who participate in curtailment programs incur *adaptation costs* and *curtailment costs*.

Adaptation costs are costs that the customer incurs just to be prepared for curtailments, regardless of whether any curtailments actually occur. For example, the customer might purchase back-up generators and devices to protect equipment from curtailment-related damage. As another example, the customer might develop special operating procedures just to be ready to deal with curtailments.

Curtailment costs are incurred when customers' loads are actually curtailed. Depending upon the type of customer, there may be lost or damaged production, production re-start costs, lost sales, food spoilage, and inconvenience.

From the perspective of customers, curtailment program costs are highest when the program has precisely the same characteristics as those that make curtailment values highest for the power system.

6.1.3. Quantifying curtailment credits for the traditional design

LSEs should be willing to pay to customers no more than the customer's participation in the curtailment program is worth to the LSE. Over a period like a year, the LSE's payments (or credits) to the customer should not exceed the sum of: a) the expected insurance value of the non-spinning reserves provided by the customer; plus b) the expected operating value (or net energy cost savings) of actual curtailments. The maximum payment or credit in any year Y to any particular customer can be expressed as:

$$PMT_Y^1 \leq \sum_{h \in Y} E\left\{Q_h^{AV} * \left(P_h^{NSR} - C^{AV}\right) + Q_h^{CURT} * \max\left[0, P_h^{EN} - P_h^{RET} - C^{CURT}\right]\right\} - C_Y^{FIX} \quad (6)$$

where the variables are as follows:

PMT_Y^1	the annual payment or credit for the traditional rate design ($)
E	the expectation operator
Q_h^{AV}	the load that is available to be curtailed in hour h (MWh)
P_h^{NSR}	the forward wholesale market price of non-spinning reserves in hour h ($/MW/hr)
C^{AV}	program costs that depend upon the quantity of curtailable load ($/MW/hr)
Q_h^{CURT}	the load that is actually curtailed in hour h (MWh)

P_h^{EN} the forward wholesale market price of energy in hour h ($/MWh)

P_h^{RET} the retail price of energy in hour h ($/MWh)

C^{CURT} program costs that depend upon the quantity of curtailed load ($/MW/hr)

C_Y^{FIX} annual program costs that depend upon the numbers of participants ($).

The right-hand side of equation 6 indicates that the insurance value of curtailable load in any hour is equal to the quantity of curtailable load Q_h^{AV} times the wholesale price of non-spinning reserves P_h^{NSR} net of certain program costs (C^{AV}), where "forward wholesale market price" refers to day-ahead or hour-ahead prices.. The operating value of curtailed load equals the quantity of curtailed load Q_h^{CURT} times the excess of the wholesale price of energy P_h^{EN} over the retail energy price P_h^{RET} and certain program costs (C^{CURT}).

6.1.4. Quantifying credits and payments for the performance-based design

Again, LSEs should be willing to pay to customers no more than the customer's participation in the curtailment program is worth to the LSE. And again, over a period such as a year, the LSE's payments (or credits) to the customer should not exceed the sum of: a) the expected insurance value of the non-spinning reserves provided by the customer; plus b) the expected operating value of actual curtailments. But, unlike the traditional design, the performance-based design has separate payments for the insurance value and the operating value. Furthermore, because the payments for operating value occur soon after curtailment, they can be based upon observed day-ahead or hour-ahead market energy prices rather than year-ahead or season-ahead forecast prices.

The maximum payments or credits in any year Y to any particular customer can be expressed as:

$$PMT_Y^{INS} \leq \sum_{h \in Y} E\left\{Q_h^{AV} * \left(P_h^{NSR} - C^{AV}\right)\right\} - C_Y^{FIX} \quad (7a)$$

$$PMT_Y^{OP} \leq \sum_{h \in Y} Q_h^{CURT} * \max\left[0, P_h^{EN} - P_h^{RET} - C^{CURT}\right] \quad (7b)$$

where

PMT_Y^{INS} the annual payment or credit for program participation under the performance-based design

PMT_Y^{OP} the payment or credit for actual curtailments under the performance-based rate design

and the remaining variables have the same meanings as for the traditional design (though not necessarily the same values). Equation 7a represents the payment for the insurance value of curtailable load. Equation 7b represents the payments for the operating value of actual load curtailments.

6.1.5. Comparison of the traditional and performance-based designs

The differences between equations 6 and 7 are more significant than they might appear. On the surface, the only differences between the equations seem to be: a) the performance-based design has separate payments for the insurance and operating values; and b) the payment for operating value is based on expected curtailments and prices under the traditional design, while it is based on actual curtailments and prices under the performance-based design. The differences are more significant than that, however, because the two program designs give different incentives to LSEs and customers, and they also impose different risks on LSEs and customers. Consequently, although the variables in equations 6 and 7 share identical notation, they can take on very different values.

Under the traditional design, payments to the customer for actual curtailments are divorced from the customer's direct experience of curtailments. Since the customer gets the payment regardless of whether curtailment actually occurs, the customer has a strong incentive to avoid actual curtailments because they are purely a losing proposition. Therefore, customers' cooperation with curtailment calls can be grudging or worse. Furthermore, the traditional design exposes customers to substantial risk: while the customer may have joined the curtailment program in the good-faith belief that actual curtailments would be significantly less frequent and severe than the maxima allowed by contract, the customer may instead be unpleasantly surprised to find that extreme power system conditions lead the LSE (or the system operator) to exercise its full curtailment rights.

The traditional design also creates risks for the LSE. The LSE's payments for curtailable load reflect expectations of what actual curtailments will be. Mild system conditions will lead the LSE to derive less value from its curtailable loads than it expected, meaning that it paid too much for a service that it did not need. In principle, one might expect LSEs to overpay for curtailable service in some years and to underpay in other years, thus paying a fair price in the long run. As a matter of historical fact, however, the vast majority of utilities have consistently overpaid for curtailable service because their relationships with customers are happier when customers are overpaid than when customers are underpaid.

The performance-based design directly matches payments for curtailments with actual curtailment volumes and day-ahead or hour-ahead energy prices. Although customers may not like being curtailed, they receive payments every time a curtailment occurs, with larger payments for longer or more severe curtailments. These payments are likely to make customers more cooperative in accepting curtailments. Furthermore, the more curtailments a customer experiences, the more payments that it will receive; so the risk of an unpleasant surprise of numerous curtailments is at least partly balanced by the receipt of more numerous payments.

For the LSE, the performance-based design assures that the LSE pays for the curtailments that it gets, and not for curtailments that never happen.

In summary, traditional programs have the relative advantage of substantial precedence, which makes them familiar and acceptable to customers. Because utilities have historically overpaid for curtailable loads that they have not actually curtailed, traditional programs have also had a history of satisfactory participation rates by customers.

Performance-based programs have the relative advantage that they are likely to lead to more efficient use of curtailments by LSEs, more cooperative and efficient responses by customers to curtailment calls, and lower financial risks to both the LSE and its customers. In short,

performance-based programs allow LSEs to pay customers for the curtailment services that customers actually deliver.

Examples of the performance-based design are provided by reliability-based demand response programs offered by ISO New England and the New York ISO.[31] Under both programs, curtailments are mandatory, and participating customers receive both a capacity payment (analogous to a discounted demand charge) and a payment for actual curtailments (at a guaranteed minimum price tied to the real-time market price) when called by the ISO.

6.2. Information Requirements

The information required to determine curtailable service credits and payments is defined by the variables that appear in equations 6 and 7. These variables are the quantities of curtailable and curtailed load, and the expected wholesale prices of energy and non-spinning operating reserves.

6.2.1. Quantifying curtailable load and curtailed load

Curtailable load Q_h^{AV} and curtailed load Q_h^{CURT} need to be forecast according to the quantities of curtailment that the utility can actually expect. These quantities depend upon:

- *Customers' hourly loads*. The amount of curtailment in any hour can never exceed the customer's load in that hour and is generally limited to the excess of the customer's load over their firm power level (if any). If the customer's load happens to be low before a curtailment occurs, the customer may be unable to reduce load by very much. Forecasts of curtailable load should anticipate the variability of customers' non-curtailed loads.
- *Curtailment dispatch procedures.* Because there are contractual limits on the frequency and duration of curtailment calls, curtailments will not necessarily be called in the hours when curtailments are most valuable. The estimated quantities need to anticipate that dispatch will sometimes "miss" picking the most valuable hours.
- *Customer cooperation.* Different customers have different ability and willingness to cooperate with curtailment calls. Forecasts of curtailable load should reflect individual customers' history of cooperation (or non-cooperation) with curtailment calls, or should be based upon histories of similarly situated customers.

Although contractual terms regarding the customer's curtailment obligations indicate what the curtailment quantities might be, they do not necessarily indicate what curtailable load or curtailed load will actually be.

6.2.2. Forecasting energy and reserve prices

Prices can be forecast by the methods described in Section 5.4.2.

[31] The ISO New England and NYISO websites provide descriptions of their demand response programs.

6.2.3. Quantifying program costs

Program costs can be derived, fairly accurately, from historical experience with existing curtailable service programs. A particular challenge, however, will be dividing those costs into the three categories: those that depend upon the quantity of curtailable load (C^{AV}); those that depend upon the quantity of curtailed load (C^{CURT}); and those that depend upon the numbers of participants (C_Y^{FIX}). This division may be achieved through some combination of cost analysis and statistical analysis.

7.0 Phase II Plan

7.1. Research Activities

The overriding objective of the Phase II project is to demonstrate the development a set of retail rates that provide appropriate incentives for efficiency and demand response and are acceptable to the various stakeholders in the process, including utilities, regulators, and consumer representatives. These rates can take the form of default (opt out) or voluntary (opt in) rates, but will have to be designed in the context of existing utility rate cases and CPUC activities in the dynamic pricing area. The background of strong interest in such rates in California, combined with failure to agree on a rate design that satisfies all parties, suggests the need for a collaborative process in which the technical consultant on this project serves as an independent party to solicit and take into account the concerns of the stakeholders, conduct the needed rate design analysis, and demonstrate that the selected rates can satisfy the demands of the various parties.

Successfully achieving the project objectives will require several different types of activities, including *review and analysis* of existing regulatory barriers to adoption of demand-responsive pricing alternatives, development of a set of *evaluation criteria* for rating potential pricing alternatives, *technical analysis* to develop specific pricing elements of the rates, and *communication and coordination* with project stakeholders and other project consultants.

In high-level terms, we envision the Phase II project as consisting of the following main elements:

1. Assessing the current status of rate regulation practices in California that stakeholders view as limiting the incentives or ability of utilities to implement more efficient rates. An early step in the project will involve a *review of documents*, including existing retail rates, recently proposed CPP and RTP rates, and descriptions of specific regulatory policies such as revenue recovery mechanisms. Another step will involve *interviews with relevant staff* at utilities and regulatory agencies to identify key regulatory constraints on retail pricing designs. We will also solicit a case study utility to work with on the Phase II project, and to provide needed customer load data.
2. Identifying a list of candidate rate options that meet the stakeholders' interest in providing incentives for efficiency and demand response and that have the potential to address the regulatory issues identified in the first step. The list may include, for example: a) RTP with hedging and a performance-based curtailable load program for large C&I customers; b) CPP built on a TOU rate structure for medium C&I customers; and c) CPP built on existing rate structures for residential and small C&I customers. The candidate list will be developed through a scoring procedure based upon a set of evaluation criteria consisting partly of the conceptual framework established in this Phase I report (*i.e.*, the economic efficiency criterion that retail energy prices reflect expected marginal costs by time period and location), and partly on factors such as degree of program complexity and likely customer acceptance. The list of candidate rate options will be presented to the stakeholders at a meeting designed to reach consensus on the rate options to be explored further, and to discuss methods for obtaining market cost and customer load data needed to calculate specific pricing elements.

3. Developing examples of specific pricing elements (*e.g.*, TOU prices, critical peak prices, and formulas for calculating RTP prices) for the candidate retail rate options. These elements will be based on the pricing formulas presented in Section 5.2 of this report, and on data on costs and customer loads, where these data are developed possibly in conjunction with other DRRC demand response and rate projects.
4. Comparing the fully specified candidate rate options to the relevant standard tariffs to assess the feasibility of offering the rate options as optional or default tariffs.

The Phase II project will involve extensive communication and interaction with the stakeholders, including staff from the utilities, regulatory agencies, and consumer representatives. As requested in the RON, we will work with DRRC to establish a communication and outreach plan to ensure that we stay abreast of ongoing dynamic pricing activities at the utilities, the CPUC and the CEC. We will communicate research results to key stakeholders at periodic meetings.

We anticipate that the project will involve a series of iterative steps, with at least three project meetings to review project objectives and accomplishments, and to confirm the direction of the next step. The *first meeting* will be designed to agree on project objectives, conduct a structured discussion of existing rate regulation features that limit the incentives and ability of California utilities to develop demand-responsive rates, review candidate pricing options, solicit a case study utility, and discuss sources of marginal cost and load data. This meeting could be expanded to represent a workshop including each project team working on the DR valuation and rate design projects.

The *second meeting* will involve a review of preliminary versions of selected rate options, including specific pricing elements calculated from the available marginal cost and load data. Information will also be provided on estimates of potential cost-saving *benefits* of the rate options to consumers and the utility, estimates of potential *market acceptance* based on those benefits relative to the current standard tariffs, and an evaluation of the *feasibility* of the rate options given the regulatory issues identified at the first meeting and suggestions for changes needed to make the rate options more feasible.

The third meeting will present final versions of the recommended rate options, as modified following discussions at the second meeting. The meeting will also discuss next steps for possible roll-out and marketing of one or more of the rate options analyzed in the project.

7.2. Project tasks

To achieve the overall objectives, Phase II will involve a series of tasks, all of which will be conducted in close collaboration with the technical advisory group, the utilities, the CEC, and the CPUC. These tasks are described below.

7.2.1. Task 1. Identify issues and objectives, candidate rate structures, and case study utility.

This task will involve review and development of background material based on discussions with project advisors and stakeholders, and will involve a project meeting to reach agreement on project objectives, a preliminary list of retail rate structures, and a case study utility. The task will include the following activities:

1. *Identify key issues and objectives*. Prior to the first project meeting, we will interview key stakeholders to identify short-term and longer-term rate regulation issues that appear to present barriers to utility acceptance of dynamic pricing products that promise to provide efficient demand response. The initial meeting will be designed to reach a consensus on the primary objectives of the project, the nature of the rate regulation features that have presented barriers to more efficient pricing, and a list of potential retail rates to be designed and evaluated in the project.
2. *Identify case study*. We will solicit at least one of the California utilities to serve as a case study for developing specific examples of dynamic pricing for one or more customer classes. The utility would agree to provide customer load data for samples of customers in the relevant classes. We believe that involvement by at least one of the major utilities is crucial to the ultimate success of the Phase II project, and will provide needed input on practical considerations in designing and implementing new retail rate structures.
3. *Identify candidate rate structures*. We will develop a list of candidate rate structures that could be offered to each customer class, and that could be developed into specific rates based on actual marginal cost and customer data. We will conduct a preliminary evaluation of the candidate rate structures relative to a set of criteria based on our conceptual framework (*e.g.*, the extent to which the rate is likely to reflect differences in wholesale market costs by time period and location), as well as factors such as degree of complexity and likely customer acceptance.

Deliverable. Memorandum providing background material on key project issues and candidate rate structures. The memorandum will serve as discussion material for a project kick-off meeting to establish project objectives, agree on a list of candidate rate designs, and select a case study utility.

7.2.2. Task 2. Develop preliminary versions of selected rate structures.

In this task we will develop preliminary estimates of the specific pricing elements of one or more rate options selected in the first project meeting, and produce estimates of the potential benefits of that rate option to participating customers and the utility. The pricing elements will be based on applying the pricing formulas described in Section 5.2. These require information on expected marginal costs and customer loads. Thus, a key activity in this task will be the development of these data.

1. Two key inputs to designing retail tariffs that are based on marginal costs are appropriate scenarios of expected hourly marginal costs (or market prices, if available and appropriate) and hourly loads for the customers to be served by the rates. In this task we will develop one or more scenarios of marginal costs or wholesale market prices that interested parties agree represent reasonable forecasts of conditions in the state in the short term. It would be useful and efficient if these wholesale market scenarios were developed in conjunction with other Phase II DR and rate projects.
2. We will also develop load data for the customer classes for which the selected retail rates will be designed. We will work with the case study utility to select the appropriate load data for use in the study. Data should be readily available for customer accounts of size greater than 200 kW. For those customers, a representative sample of customer accounts

will be most practical and appropriate. For mass market customers, data from load research samples may be used.

3. As input to the process of evaluating potential menus of pricing products, we will also develop assumptions about customer price responsiveness by customer type, and about the market share of each pricing product. The price response assumptions will be based on the extensive literature available on estimates of customer responsiveness (*e.g.*, price elasticities) to time-varying prices such as TOU, CPP, and RTP, including the recent evaluation of the Statewide Pricing Pilot. Estimates of market share can be informed by studies of customer perceptions and preferences regarding the features of dynamic pricing and demand response programs, including studies conducted in conjunction with the SPP, evaluations of the California demand response programs, an evaluation of the NYISO demand response programs, and recent DRRC/LBL reports on real-time pricing in New York and elsewhere.[32]

4. The preliminary versions of the rate structures selected for analysis in this task will be presented at a project meeting. The objective of the meeting will be to give the stakeholders a mid-course review of the rate structures, the methods used to calculate the pricing elements, and the specific pricing elements calculated in this task. The preliminary rates will be evaluated in terms of the incentives provided for appropriate demand response; the potential benefits of the rates to participants, the utility, and the state; potential customer acceptance; and the extent to which modifications to existing rate regulation provisions may be needed to make the rate effective and acceptable.

Deliverable. Memorandum providing description of the data and methods used in developing the preliminary versions of the selected rate structures, and summarizing the features of the rates and the estimated benefits to consumers and the utility.

Task 3. Develop final versions of selected rate structures.

Based on comments and suggestions received from the second project meeting, we will proceed to develop final versions of each of the rate structures recommended by project stakeholders. The final versions of the rate structures will be developed using similar methods as in Task 2. We will also calculate the potential benefits to the various parties for each rate, and evaluate potential customer acceptance. Final versions of the rates will be presented at a final project meeting and in a final report.

Deliverable: Memorandum describing final versions of the selected rate structures.

Task 4. Meetings.

This task will cover the preparation for and participation in the three project meetings described in Tasks 1 – 3.

Deliverables. Meeting materials, participation in three meetings, and follow-up meeting notes.

---

[32] See Barbose *et al* [2004], Neenan *et al* [2003], and Quantum Consulting [2004].

Task 5. Final report.

This task will cover preparation of a final report documenting the methods, data, recommendations, and final versions of the rates developed in the project.

Deliverable. Final report.

7.3. Timeline

We anticipate that the Phase II project can be implemented over the remainder of 2006. This timing could in principle allow implementation of one or more of the recommended rates by the summer of 2007. Assuming that Phase II decisions and contracting details can be worked out by April, the following timeline and milestones can be achieved.

- Kickoff meeting and project planning—May 2006.
- Case study selection, data development, and preliminary rate designs—September 2006.
- Development of final versions of selected rates—November 2006.
- Final report—December 2006.

8.0 Conclusions and Recommendations

The implications of the conceptual framework presented in this report are straightforward. First, retail rates that provide appropriate incentives for efficiency and demand response are those whose prices reflect the time-varying and location-specific nature of the marginal costs of generating and delivering electricity to specific groups of consumers. Second, a variety of time-varying rate structures may be designed, each of which reflects expected marginal costs with different degrees of accuracy, price guarantees, and notice of price changes. Financial hedging mechanisms may also be used to limit consumers' risk due to price uncertainty under certain variable-price products such as real-time pricing. Third, if consumers are given a limited menu of rate structures that are characterized by different degrees of price uncertainty and that provide consistent risk premia, then consumers will, in principle, willingly provide efficient levels of demand response.

Finally, there are a number of practical and transitional issues that are likely to hinder development of efficient demand response rates, as California has already seen over the past few years. In particular, existing ratemaking practices may impede the efficient design of both standard and time-varying retail rates. However, we strongly believe that a consistent application of the fundamental conceptual framework provided in this report can provide a roadmap to the successful development of efficient, demand-responsive retail rates. After all, we are only trying to implement through regulated rates what competitive market forces would achieve in a workably competitive energy market in the absence of regulation. Furthermore, previous investments in advanced metering equipment have provided a ready infrastructure to support demand responsive retail rates for a large fraction of the electricity load in the state.

A high priority should be given to exploring and addressing the effects and limitations of existing regulatory practices and practical issues in follow-on research and development in Phase II.

References

C. Aubin, D. Fougere, E. Husson, and M. Ivaldi, "Real-Time Pricing of Electricity for Residential Customers: Econometric Analysis of an Experiment," *Journal of Applied Econometrics*, Vol. 10, S171-S191, 1995.

M.L. Baughman and S.N. Siddiqi, "Real-Time Pricing of Reactive Power: Theory and Case Study Results." *IEEE Transactions on Power Systems*, 6(1): 23-29, February 1991.

G. Barbose, C. Goldman and B. Neenan, "A Survey of Utility Experience with Real Time Pricing," LBNL-54238, December 2004.

A.W. Berger and F.C. Schweppe, "Real Time Pricing to Assist in Load Frequency Control," *IEEE Transactions on Power Systems*, 4(3): 920-26, August 1989.

R.E. Bohn, M.C. Caramanis, and F.C. Schweppe, "Optimal Pricing in Electrical Networks Over Space and Time," *Rand Journal of Economics*, 15(3): 360-76, Autumn 1984.

M. Boiteux, "La Tarification des Demandes en Pointe," *Revue Generale de l'Electricite*, 58: 321-40, 1949. Translated as "Peak Load Pricing", *Journal of Business*, 33(2): 157-79, 1960.

J.C. Bonbright, A.L. Danielsen, and D.R. Kamerschen, *Principles of Public Utility Rates*, Public Utilities Reports, Inc., 1988.

S. Borenstein, "The Long-Run Efficiency of Real-Time Electricity Pricing," *The Energy Journal*, Vol. 26, No. 3, 2005.

S. Braithwait, "Residential TOU Price Response in the Presence of Interactive Communications Equipment," in *Pricing in Competitive Electricity Markets*, edited by A. Faruqui and K. Eakin, Kluwer Academic Publishers, 2000.

S. Braithwait and D. Armstrong, "Potential Impact of Real-Time Pricing in California," Report to California Energy Commission, January, 14, 2004.

S.D. Braithwait and D.W. Caves, "Three Biases in Cost-Efficiency Tests of Utility Energy Efficiency Programs," *The Energy Journal*, Vol. 15, No. 1, 1994.

S. Braithwait and M. O'Sheasy, "RTP Customer Demand Response–Empirical Evidence on How Much Can You Expect," *Electricity Pricing in Transition*, edited by A. Faruqui and K. Eakin, Kluwer Academic Publishers, Boston, MA., 2002.

G. Brown, Jr. and M.B. Johnson, "Public Utility Pricing and Output Under Risk," *American Economic Review*, March 1969.

S.J. Brown and D.S. Sibley, *The Theory of Public Utility Pricing*, Cambridge University Press, 1986.

California Energy Commission, *2005 Integrated Energy Policy Report*, November 2005.

California Public Utilities Commission, *Decision Closing This Rulemaking and Identifying Future Activities Related to Demand Response*, Rulemaking 02-06-001 (Order Instituting Rulemaking on policies and practices for advanced metering, demand response, and dynamic pricing), November 18, 2005.

M.C. Caramanis, R.E. Bohn, and F.C. Schweppe, "System Security Control and Optimal Pricing of Electricity," *Electric Power and Energy Systems*, 9(4): 217-24, October 1987.

H. Chao, "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty," *Bell Journal of Economics*, Spring 1983.

Charles River Associates, "Impact Evaluation of the California Statewide Pricing Pilot," March 16, 2005.

Christensen Associates Energy Consulting, "Evaluation of California's Real-Time Energy Metering (RTEM) Program," Final report to the California Energy Commission, March 7, 2005.

M.A. Crew and P.R. Kleindorfer, "Peak Load Pricing with a Diverse Technology," *Bell Journal of Economics*, Spring 1976.

A. Faruqui and J.R. Malko, "The Residential Demand for Electricity by Time of Use: A Survey of Twelve Experiments with Peak Load Pricing," *Energy* 8(10):781-795, 1983.

J.D. Glyer, "The Price Signal of "Flip-the-Switch" Products: You Can Run but You Can't Hide," EPRI International Energy Pricing Conference, Washington, DC, 2000.

W. W. Hogan, "On An 'Energy Only' Electricity Market Design for Resource Adequacy," John F. Kennedy School of Government, Harvard University, September 23, 2005.

A. Jenkins, "Real-Time Pricing Alive and Well in Georgia," *Energy Pulse*, March 17, 2005.

A.E. Kahn, *The Economics of Regulation, Principles and Institutions*, The MIT Press, 1988 (first published as two volumes in 1970-71 by John Wiley & Sons, Inc.).

J.A. Kay, "Uncertainty, Congestion, and Peak-Load Pricing," *Review of Economic Studies*, October 1979.

B.M. Mitchell, W.G. Manning, Jr. and J.P. Acton, *Peak-Load Pricing, European Lessons for U.S. Energy Policy*, Ballinger Publishing Company, 1978.

B. Neenan, D. Pratt, C. Goldman, G. Barbose, *et. al*, "How and Why Customers Respond to Electricity Price Variability: A Study of NYISO and NYSERDA 2002 PRL Program Performance," Report to NYISO and NYSERDA, January 2003.

S.S. Oren, "Generation Adequacy via Call Options: Safe Passage to the Promised Land," *The Electricity Journal*, November 8, 2005.

J.C. Panzar, "A Neoclassical Approach to Peak Load Pricing," *Bell Journal of Economics*, Autumn 1976.

Quantum Consulting, Inc. and Summit Blue Consulting, "Working Group 2 Demand Response Program Evaluation – Program Year 2004," December 2004.

W. Samuelson and R. Zeckhauser, "Status Quo Bias in Decision Making," *Journal of Risk and Uncertainty*, Springer, Vol. 1(1), pages 7-59, 1988.

B. Schwartz, *The Paradox of Choice*, HarperCollins Publishers, Inc., New York, NY, 2004

Southern California Edison, *Phase 2 of 2006 General Rate Case Marginal Cost And Sales Forecast Proposals*, before the Public Utilities Commission of the State of California, Application No. 05-05-023, Exhibit SCE-2 (Updated), September 6, 2005.

Southern California Edison, "Testimony of Southern California Edison Company Proposing Default Critical Peak Pricing Rate Design for Large Customers," before the Public Utilities Commission of the State of California, August 1, 2005.

P.O. Steiner, "Peak Loads and Efficient Pricing," *Quarterly Journal of Economics*, November 1957.

N.D. Uri, "Peak Load Pricing with Spatial Diffusion," *Regional Science and Urban Economics*, 6(2): 161-72, May 1976.

M.L. Weitzman, "Prices vs. Quantities," *Review of Economic Studies*, 41(4), No. 28, pp. 477-91, October 1974.

R.B. Wilson, *Nonlinear Pricing*, Oxford University Press, 1993.

Appendix A.
Theory of Public Utility Pricing

This appendix provides a brief history of the theory of public utility pricing. It is designed to provide an historical context for the conceptual framework for efficient retail rates developed in this report.

An updated version of the often-cited bible of public utility rate design, Bonbright, Danielsen, and Kamerschen [1988] lists a series of attributes of a sound rate structure. Grouped into categories of *revenue-related*, *cost-related*, and *practicality*, they include the following:

Revenue-related

1. Effectiveness in yielding a utility's revenue requirement
2. Revenue and rate stability and predictability

Cost-related

3. Static efficiency in discouraging wasteful use of service, while promoting all justified types and amounts of use, including relative uses of types of service (*e.g.*, on-peak versus off-peak service)
4. Reflection of all current and future private and social costs and benefits of providing service
5. Fairness in apportioning total costs of service among different ratepayers (so as to avoid arbitrariness and capriciousness)
6. Avoidance of undue discrimination and cross-subsidies
7. Dynamic efficiency in responding to changing demand and supply patterns

Practicality

8. Simplicity, understandability, and public acceptability.

These attributes derive from three fundamental objectives of public utility regulation:

- *Capital attraction*. Utilities need sufficient revenue to recover costs and earn a fair return that will allow them to attract capital needed to invest in equipment for supplying its customers.
- *Consumer rationing*. Rates should be designed to support economically efficient usage, that is, discouraging wasteful use of public utility services, while promoting "all use that is economically justified in view of the relationships between the private and social costs incurred and benefits received."
- *Fairness to ratepayers*. The revenue required by utilities should be recovered from customers through cost allocations that are not arbitrary or capricious.

A Brief History of Marginal Cost Pricing Theory

Since the late 1800s, economists have recognized that prices are most efficient when they equal marginal costs, and that the "invisible hand" will automatically lead prices to approximate marginal costs in a truly competitive market. "Marginal costs" are defined as the change in all costs that accompany a change in the output of a good. For example, the marginal cost of providing 1 MWh of electrical energy to a consumer would include all the extra costs that the power system, customers, and non-customers incur to provide that extra MWh. These costs include fuel, labor, financial risk, and environmental costs, wherever and whenever they may have been incurred to get that extra MWh to the customer.

The mathematics of the marginal cost pricing of electricity were first published (in French) by Boiteux [1949]. The seminal English language article was that of Steiner [1957], which examined a power system with a single generator. Steiner found that off-peak customers should pay a price that only reflects marginal operating costs, that on-peak customers should pay a higher price that reflects both marginal operating costs and marginal capacity costs, and that optimal generation capacity levels are determined by on-peak customers' willingness to pay for capacity.

Since Steiner, economists have added numerous insights to his basic ideas concerning the relationships between marginal costs, prices, and capacity. First, economists developed models in which power systems had many generators, not merely one.[33] They found that, for each time period that has a different level of demand, there will be a different efficient price determined by the price level at which the demand curve intersects the marginal cost curve. Consequently, one implication of power systems having many generators is that there can be many efficient price levels: the efficient market price can be different in each time period because there can be a different marginal generator with a different marginal cost in each time period. Because a generator will recover part of its capital costs whenever the market price exceeds its marginal operating cost, which occurs whenever a more expensive generator is needed to meet market demand, generators can recover parts of their costs in many different periods, not just in the peak period.

Second, economists considered the effects of uncertainty upon marginal costs.[34] These uncertainties arise because of load fluctuations and because of generator and transmission facility outages. Consequently, the identities of marginal generators cannot be forecast with certainty, nor can marginal costs be forecast with certainty. Prices should reflect the costs that the power systems and customers incur (or might incur) to either avoid or deal with these uncertainties.

Third, economists (and their engineering brethren) recognized that the marginal costs of delivering power to customers vary according to the customers' locations.[35] These spatial variations occur because of transmission losses and constraints. The mathematics thus developed serve as the basis for present-day calculations of locational marginal prices (LMPs).

---

[33] See Crew and Kleindorfer [1976] and Panzar [1976].

[34] See Brown and Johnson [1969], Chao [1983], and Kay [1979].

[35] See Bohn, Caramanis, and Schweppe [1984] and Uri [1976].

Fourth, as power system unbundling has progressed, economists have realized that there are different marginal costs for each of the unbundled services. Methods have thus been proposed for estimating the marginal costs of regulation, operating reserves, and reactive power.[36]

Dr. Alfred Kahn, who once served as Chairman of the New York Public Service Commission, explained that the fundamental rationale for setting retail prices so as to signal marginal cost is to achieve an efficient allocation of society's limited resources.

> "At any given time, every economy has a fixed bundle of productive resources, a finite total potential productive capacity. Of course, that total can grow over time; but at any given time the basic economic problem is to make the best or most efficient use of that limited capacity. The basic economic problem, in short, is the problem of choice. A decision to produce more of any one good or service is, in these circumstances, ipso facto a decision to produce less of all other goods and services taken as a bunch. It follows that the cost to society of producing anything consist, really, in the other things that must be sacrificed in order to produce it; in the last analysis, "cost" is opportunity cost—the alternatives that must be foregone. In our economy, we leave the final decision about what shall be produced and what not to the voluntary decisions of purchasers, guided by prices on the one hand and their own wants or preferences on the other."[37]

As characterized by Kahn, marginal cost "must include all the costs that production of an additional unit imposes, regardless of when those costs are actually realized." As emphasized by Kahn and Bonbright *et al*, the crucial role of marginal cost relates to the concept of *opportunity cost*—"the value of anything is calibrated in terms of lost alternatives or opportunities—because all resources are limited."

A Brief History of Electricity Pricing

1950s and 1960s

Despite the long recognition of the principle of sending economically efficient price signals to customers that reflect the marginal cost of generating and delivering electricity, the practice of incorporating marginal cost-based pricing in utility rate making has been slow in implementation. The period of the 1950s through the early 1970s was characterized generally by declining electricity costs, which provided little urgency for pricing reform. With marginal costs less than average costs, there was for the most part only academic interest in efficient, marginal cost-based pricing, with concern expressed that pricing at marginal cost would not recover sufficient revenue to cover utilities' total costs.

The 1970s Energy Crises

The situation changed dramatically during and after the two energy crises of 1973-74 and 1979-80. A series of factors combined to produce marginal costs that exceeded average costs by a

---

[36] See Baughman and Siddiqi [1991], Berger and Schweppe [1989], and Caramanis, Bohn, and Schweppe [1987].

[37] Kahn [1988].

considerable amount. These included higher fossil fuel prices, double-digit inflation and interest rates, rising construction costs, particularly for nuclear plants, and an apparent end to economies of scale in power plant construction.

The increased pressure on electricity prices led to new interest in setting retail prices to reflect marginal costs, and to thus encourage more efficient use of energy and less need for costly new capacity. This interest led to the quickened development of academic theory on electricity marginal costs, cited above. It also encouraged the National Association of Regulatory Utility Commissioners (NARUC), the Edison Electric Institute, and EPRI to collaborate in drafting a comprehensive *Electric Utility Rate Design Study* to assess appropriate methods for estimating marginal costs and the "feasibility and cost of shifting various types of usage from peak to off-peak periods." This multi-year study resulted in multiple volumes of reports on marginal costs, time-of-use pricing, load management, and related topics.[38]

At around the same time, the Federal Energy Administration, the predecessor of the Department of Energy, funded a series of residential time-of-use (TOU) pricing experiments to assess how customers would respond to peak and off-peak prices. Mitchell, Manning, and Acton [1978] reported that by 1977, state rate proceedings in California, Michigan, New York and Wisconsin led to the first use of TOU rates for very large customers. For example, a Pacific Gas and Electric (PG&E) tariff moved large customers from old declining-block demand and energy charges to new seasonal demand and energy prices that differed by peak, shoulder, and off-peak periods.

Not long after, Congress passed the Public Utility Regulatory Policy Act (PURPA) in 1978, which required state regulatory commissions to consider the cost effectiveness of retail rates designed to reflect differential costs of serving various types of customers, and the variation of costs by time of day and season, as well as provision of interruptible service rates for large industrial consumers. PURPA led to considerable expansion of load research metering to establish the costs of serving different customer classes, and to more widespread and explicit use of marginal costs in retail rate design. Interestingly, the recently passed Energy Policy Act of 2005 modifies various sections of PURPA to further require utilities and state commissions to assess the potential value of advanced metering equipment and of time-based rate and demand response programs.

Late 1980s and 1990s

The global movement toward restructuring and/or deregulation of regulated industries such as railroads, airlines, and telecommunications also encompassed the electric power industry, producing new trends in electricity pricing. A series of EPRI pricing conferences during the 1990s carried themes such as "Pricing in Competitive Electricity Markets," "Market-Based Pricing," and "Pricing in Transition." Utilities saw competition in various forms, including competition from other fuel types such as natural gas, competition for large customers from other regions within or outside of the state, and, in states that allowed retail competition, for all customers.

[38] See Faruqui and Malko [1981] for a summary of the TOU pricing studies.

One effect of competition and industry restructuring was cost unbundling into customer, wires, and energy services. In addition, one of the pricing mechanisms used to address competition was real-time pricing (RTP), in which hourly prices are based on anticipated marginal costs, typically for the following day. PG&E is generally credited with starting the first RTP program in the U.S. in the mid 1980s. However, the prices in that program far exceeded marginal costs because of that program's *one-part* design: all RTP revenue requirements were recovered solely through the RTP energy charge. To get efficient energy prices, Niagara Mohawk instituted, in 1988, the first *two-part* RTP design, in which hourly prices reflected hourly marginal costs and revenue recovery was achieved through access charges on a baseline level of usage. This was followed in 1992 by a similar two-part design at Georgia Power Company, which has since grown into the largest program in the U.S.

Post 2000

A series of factors have combined in recent years to intensify interest in dynamic retail pricing that reflects marginal costs and demand response programs that encourage peak load reductions through payments that reflect marginal costs or market prices. These include the following:

- A series of extreme wholesale price spikes in 1999 in the Midwest and Eastern U.S., in which prices in a few hours rose to levels of around $8,000 per MWh.
- The California crisis of 2000-2001, in which wholesale prices repeatedly rose to their capped level, leading to financial crises and rolling blackouts.
- A recognition that organized wholesale energy markets (*e.g.*, PJM, New York ISO, ISO New England, and California ISO) need some degree of responsive demand to operate efficiently.

Much of the discussion of responsive demand in organized wholesale markets has taken place in the three regions in the Eastern U.S. in which retail competition has been initiated. Since the distribution divisions of the remaining utilities provide only default service to consumers who have not selected a competitive energy provider, there has been relatively little focus on innovative rate design on the part of those utilities. In fact, some states, such as New Jersey, have required that default service for consumers above a certain size be priced at hourly market prices. This gives consumers an incentive to select an alternative provider who will offer a less risky product. As a result of the lack of emphasis on efficient retail pricing for regulated utilities, efforts to promote responsive demand have focused on demand response programs organized by the regional ISOs.

Appendix B.
Customers' Bills and Incentives Under Hedged RTP

Two-part real-time pricing (RTP), which has an energy charge and an access charge based upon a customer baseline load (CBL), is equivalent to RTP with a financial hedge. The two are exactly equivalent if the financial hedge is for a quantity of power equal to the CBL at the same price of the CBL. The characteristics of two-part RTP and of a financial hedge can be seen by examining the algebra that defines the consumer's bills.

Assume that a consumer faces hourly RTP prices that they hedge by obtaining a forward contract for a quantity of power, CBL_h in each hour, at a fixed price P^B. In each hour, the consumer pays (P^B * CBL_h) for power under the forward contract. If the consumer's actual load is L_h, the consumer buys the excess of load over the CBL at the real-time market price P_h^R or sells the excess of the CBL over load at the real-time market price. Over all the hours h in period T, the consumer's bill is:

$$Bill_T = \sum_{h \in T} \left[P^B * CBL_h + P_h^R * \left(L_h - CBL_h \right) \right]. \quad \text{(B1)}$$

This form of the bill shows that financially hedged RTP is *bill neutral for each customer relative to his CBL*. If he maintains his usage at the CBL level in each hour (*i.e.*, $L_h = CBL_h$), then his bill remains equal to the cost of that consumption at the forward contract price. In any hour in which his usage under RTP exceeds his CBL (*i.e.*, $L_h > CBL_h$), then he pays a *charge* equal to the incremental usage times the RTP price. Alternatively, whenever he reduces usage below the CBL level under RTP, then he receives a *credit*, valued at the RTP price.

The second form of the RTP bill may be obtained by rearranging terms in the above equation to produce the following equivalent form for the RTP bill:

$$Bill_T = \sum_{h \in T} \left[P_h^R * L_h + \left(P^B - P_h^R \right) * CBL_h \right]. \quad \text{(B2)}$$

This form of the bill shows that the customer's incentives to respond to RTP prices are not affected by the forward contract. On the right side of equation B2, the first term, which represents what the customer's bill would be if he simply purchased all consumption at RTP prices, shows that the customer's *entire load* is exposed to RTP prices. That is, any change in consumption in response to RTP prices, regardless of whether usage is above or below the CBL, has a direct impact on the bill. The second term represents a series of financial adjustments that are completely independent of the consumer's actual load L_h. These adjustments instead depend only on the CBL (which is fixed) on the extent to which actual RTP prices (which are beyond the consumer's control) differ from the forward contract price (which is also fixed). Consequently, the customer's only opportunity to affect the bill through load changes is to modify load in response to the hourly RTP prices.

Equation B2 also shows the operation of the risk management feature provided by the access charge or (equivalently) by the forward contract. When RTP prices are *high* (*e.g.*, $P^B < P_h^R$), the financial adjustment is negative, thus partially offsetting the effect of the high RTP prices on the

customer's bill. When RTP prices are *low* (*e.g.*, $P^B > P_h^R$), the financial adjustment is positive, representing a risk management payment.

Two-part RTP does differ in one substantial way from RTP with a financial hedge. Under two-part RTP, the utility sets the CBL at an historical level to insure revenue neutrality at the customer's otherwise applicable retail tariff. Under RTP with a financial hedge, by comparison, the quantity of load covered by the hedge is determined by mutual agreement between the customer and the seller of the forward contract.

Appendix C.
The Consistency of Retail Pricing Incentives

The Research Opportunity Notice (RON) for this project expressed interest in the issue of the extent to which rate designs that provide appropriate incentives for demand response are consistent with incentives for energy efficiency. That is, as emphasized in this report, rate designs that provide appropriate incentives for demand response are those whose prices vary to reflect time-varying differences in marginal power costs, thus providing incentives for consumers to *reduce* usage when prices (and costs) are high, and *increase* usage when prices (and costs) are low. Improvements in energy efficiency are generally regarded as actions taken by consumers to reduce overall energy consumption while maintaining desired energy services.

One clear pricing incentive for energy efficiency is a high retail price, or at least a high price on marginal usage decisions such as that provided by an inclining block rate. However, a high price all of the time, particularly if it does not reflect marginal cost conditions, violates the principle of economic efficiency that underlies the standard conceptual framework for efficient retail rate design.

Consider the effect that a rate design that provides appropriate incentives for demand response has on consumers' incentives for energy efficiency. The possible relationships between pricing incentives for demand response and energy efficiency include the following:

- For an end-use application such as air conditioning, whose operating hours tend to coincide with relatively high marginal costs, retail rates such as TOU and TOU combined with CPP provide price signals that give consumers an incentive to improve the efficiency of those devices, thus reducing usage during high-price periods. In this case incentives for demand response are consistent with those for energy efficiency.
- For applications such as lighting, the results may differ by customer class. For example, most residential lighting presumably takes place in off-peak hours in which TOU prices would be relatively low, and thus not provide strong incentives for energy efficiency improvements. In contrast, commercial lighting tends to correlate more closely with high peak period prices, implying that TOU or RTP prices would provide consistent incentives for demand response and energy efficiency
- Residential customers respond to TOU and CPP rates by reducing their overall energy consumption as well as peak load reductions and load shifting. For example, Braithwait [2000] found that residential customers in a CPP pilot program in New Jersey, in addition to demonstrating significant peak load reductions, reduced their overall average daily usage during the summer months by approximately five percent compared to a control group.[39] These reductions likely represent behavioral decisions to forego some cooling services. However, the pricing incentives were consistent with encouraging energy efficiency, particularly for devices such as air conditioning. Evaluation of the effects of CPP rates in the California Statewide Pilot Program again

[39] The statistical significance of the overall reduction was relatively low due to considerable variation in usage changes across customers.

focused primarily on evidence of load reduction and load shifting, but also reported evidence of small reductions in overall energy consumption.[40]

- Some reports suggest that some RTP customers, having experienced clear hourly price patterns, have made decisions to invest in energy efficiency devices to reduce usage during high-cost periods.[41]

Measuring the net economic benefits of demand response

An important issue raised by questions about the consistency of rate incentives for demand response and energy efficiency involves the appropriate criteria to use to determine the net benefits or cost-effectiveness of efforts to enhance demand response and energy efficiency. As noted in the RON, California utilities and regulatory agencies have traditionally used a set of Standard Practice Manual (SPM) cost-effectiveness tests. However, the SPM tests, which were primarily designed for evaluating the cost-effectiveness of *energy efficiency* programs, are inappropriate for evaluating the effects of innovative *pricing* programs. This is the case largely because the SPM tests do not account for the effects of customer response to price changes, such as those illustrated in Figure 4.

An example of the problems associated with attempting to use the SPM tests to evaluate dynamic pricing mechanisms is provided by a recent Southern California Edison (SCE) filing on the business case for advanced metering infrastructure (AMI).[42] SCE compared the potential cost saving benefits from dynamic pricing to the cost of the metering equipment needed to support it. In reporting its business case findings, SCE used the framework of the "All ratepayer, or Societal, perspective," which is one of the SPM benefit-cost test alternatives. In this perspective, economic *benefits* from AMI and dynamic pricing consist of utility resource and operational cost savings (such as those shown in Figure 4), and *costs* consist of AMI equipment costs and administrative costs.

SCE suggested that an additional cost should be added to the calculation. This is an estimate of consumers' "value of service loss" from their foregone energy consumption when they reduce load in response to high dynamic prices. Taken in isolation, this factor, which should be accounted for when evaluating the net benefits of demand response, may leave the impression that participating consumers suffer only losses when they face occasional high prices under dynamic pricing. However, as shown above using Figure 4 and in the discussion of the allocation of the net benefits from dynamic pricing in Section 2.4.1.1, consumers achieve net benefits from their load response due to bill reductions that presumably exceed their foregone value from their load curtailments. Otherwise, they would be unwilling to provide those curtailments. A consistent analytical framework for assessing the benefits of dynamic pricing clears up confusion of this sort.

---

[40] See Charles River Associates [2005].

[41] See Jenkins [2005].

[42] Southern California Edison [2005].

Measuring the net economic benefits of energy efficiency

During debates in the early 1990s regarding appropriate measurement of demand-side management (DSM) benefits and costs, two Christensen Associates Energy Consulting staff members, Steven Braithwait and Douglas Caves, developed and published a comprehensive DSM benefit-cost test that used traditional economic welfare analysis to account for *all* of the changes in economic benefits and costs associated with DSM programs.[43] Careful examination of the components of the test demonstrated that California's Total Resource Cost (TRC) and Ratepayer Impact Measure (RIM) tests each represented special cases of the comprehensive benefit-cost test, under different implicit assumptions about certain factors. That is, all three tests can be shown to measure the same set of benefits and costs resulting from utility DSM programs, except that the TRC and RIM tests make specific assumptions regarding the value of certain key components of the test.

Our comprehensive net economic benefit (NEB) test also added three new elements to the benefit-cost accounting. These are: 1) the loss in economic value to all consumers from reducing usage in response to *rate increases* that result from DSM cost recovery; 2) participating consumers' gain in value from expanding their energy services through the *rebound* effect; and 3) an explicit accounting for the extent of *market inefficiencies* regarding energy efficiency.

The NEB test emphasizes the two fundamental sources of potential benefits from either DSM energy efficiency programs or dynamic pricing and demand response programs: inefficiencies in the *electricity market*, and inefficiencies in the *market for energy efficiency*. In both cases, market inefficiencies are defined as persistent differences between incremental market *costs*, and market *value*, as reflected in retail prices. For example, as described in Section 2, electricity market inefficiencies generally arise from the practice of setting fixed regulated retail rates that frequently differ from varying marginal costs. These frequent differences between price and marginal cost represent lost opportunities for resource cost savings and increased consumer net value, and have led to the need for dynamic pricing or demand response programs to signal customers when energy market *costs* substantially exceed standard retail tariff *prices*.

The same price versus marginal cost inefficiencies in the electricity market can imply potential benefits from energy efficiency programs, particularly those that cause consumers to reduce consumption during high-cost periods. The other possible source of benefits from DSM energy efficiency programs consists of the potential for consumers to reduce their energy service costs through investments in energy efficiency that will cause future bill savings whose discounted present value exceeds the incremental cost of the investments. If consumers already make rational tradeoffs between those potential investment costs and energy cost savings (which in turn implies little or no market inefficiencies), then there is little opportunity for additional cost-saving energy efficiency investments whose value exceeds their cost. Alternatively, if consumers are uninformed about the benefits of energy efficiency, or are prevented from making otherwise beneficial investments by certain market structure restrictions, and if utility programs can overcome those market inefficiencies at a sufficiently low cost, then participating consumers may achieve cost savings that exceed the combination of their own incremental costs and the costs of the utility program. An extensive literature exists on debates about and suggested evidence on the extent to which such energy efficiency market inefficiencies exist.

[43] Braithwait and Caves [1994].