The Importance of an Active Demand Side in the Electricity Industry

The alternative is socialized reliability solutions, which all have difficult issues. But experience strongly suggests we won't get meaningful demand response quickly, meaning there is no avoiding them in the short term. In the long term, electricity markets will only work properly if there is an active demand side.

Hamish Fraser

Consumers buy electricity whenever they turn on a light switch. But as long as they are unwilling or unable to specify the trade-off between the *quantities* they are willing to buy in a given hour and the *prices* at which they are willing to buy, then regulated (and socialized) solutions are necessary to ensure reliability of supply. They are necessary because consumers expect the lights to come on when they throw the switch. Consumers are not used to checking the price before throwing the switch (to see whether power is scarce and prices are high). They do not expect to be confronted with huge and unexpected price increases when they open their power bills at the end of a month. They are used to consuming at an averaged price at all times. And politicians are sympathetic to all these expectations. In short, socialized reliability solutions are required when demand doesn't respond to underlying price changes, and consumers and politicians are not prepared to accept the reliability consequences that can come from a marketplace for electricity.

But socialized reliability solutions, while necessary in the short term in restructured electricity

Hamish Fraser *is a consultant at National Economic Research Associates, Inc. (NERA) in New York.*
He thanks NERA's Sally Hunt, Hethie Parmesano, Gene Meehan, Jonathan Falk, and Veronica Irastorza for their valuable input, but accepts full responsibility for the opinions and any errors contained in this article.

  PII S1040-6190(01)00249-4

markets, must be phased out as soon as possible. In other markets, prices are not averaged in this way, demand is price-responsive, and consumers individually take responsibility for their reliability of supply by choosing a price at which they would rather not consume. Electricity markets will only work properly when they behave like these other markets.

I. Reliability and Efficiency in Generation

This article describes the way *reliability of supply* was achieved in the old, vertically integrated industry model and compares it to reliability solutions in the new model of horizontal competition. The transition from the old model to the new is a troublesome one. All the places that have moved toward the new world have found this transition to be painfully slow, and in some cases it is stalled. The problem is that the transition can be extraordinarily costly. California's experience is evidence of this. California has had disastrous supply reliability consequences as a result of a flawed transition to the new market structure. The Federal Energy Regulatory Commission (FERC) believes that basic decisions about the regulatory model need to be made so that the industry can complete the transition from the traditional cost-of-service model to a model that uses markets to price commodities and services.[1] It is right. Nobody believes that a smooth transition can occur without careful management and a commitment to see it through.

This article first describes the incentives for construction of new plants and the implications for reliability and efficiency. It describes the way optimal siting of generating plants was planned and the way fuel-type decisions were made in the old vertical integration model, the distortions that existed, and how it should work under the new market institutions. In summary,

- Under the old model, utility planners were required by central planners to build enough capacity so that involuntary load shedding would only occur with a predefined (low) probability.

Adequate mechanisms for an active demand side do not yet exist.

- Under the new system, the role of markets and an active demand side should ultimately replace the central planning role, yet many issues remain.

This article then describes a vision for the transition to the new industry model, from the perspective of system reliability. It describes alternative transitional solutions, and the advantages and disadvantages of each option. In summary,

- An active demand side is critical to a functional electricity industry, but little has been done yet in this regard and much work is needed.
- The installed capacity requirements that exist in some pools are there because, in the reality of restructured electricity markets, adequate mechanisms for an active demand-side do not yet exist. They must be regarded as transitional, not permanent.

II. The Old and the New Industry Models

In order to describe these concepts, it is useful to first describe the theoretical economic underpinnings of the old and new industry models, with respect to generation planning and reliability.

A. The Old Model

1. *Generation Planning.* In generation planning under the old vertical integration model, the keyword is *quantity*. A centralized system planner decides what quantity of different types of generating capacity to build by calculating the minimum cost of meeting expected future demand. This planning is done within large-scale mathematical programs (computers) that calculate what combination of different types of plants is the least-cost option and where to site plants. **Appendix I** describes how this type of so-called "cross-over" analysis is conducted.

2. *Reliability.* Reliability under the old model is defined in terms of the probability that the lights will go out; not that they go out because a consumer turned a switch off or made arrangements to have others

PII S1040-6190(01)00249-4

do so on its behalf, but because there is insufficient capacity to meet all load at times, and the system operator must force the curtailment of some load (i.e., start rolling blackouts) so as to maintain the frequency (integrity) of the transmission system.

Of course, if system planners built exactly the amount of capacity they think is needed based on *expected* demand, shortages—and thus blackouts—would occur in 50 percent of years, or more. The reason? Load forecasts and other guesses used in a generation planning analysis are stochastic variables. With a "perfect" forecast, half the time the guess is too high and half the time the guess is too low. The guess is not expected to be exactly right every time. Because of this, if system planners just built the capacity specified by a deterministic analysis, reliability would be too low—peak load would exceed total capacity half the time. System planners have historically aimed for a probability of outage more in the order of one day in 5 or 10 years.

To ensure these types of reliability levels, systems are planned with extra installed capacity. Installed reserve margins are determined based on the target reliability levels, and using stochastic analyses of demand forecast errors, generation and transmission outages, and other factors.

B. The New Model

The notions of generation planning and reliability in the new industry model are considerably different than in the old.

1. Generation Planning. In generation planning under the new (market-based) model of industry organization, the keyword is *price.* Decentralized market participants plan generation investments on the basis of their expectation of future energy prices (and perhaps ancillary services prices and capacity prices). Setting ancillary services and capacity aside for a moment, they build new plants when energy prices are high and do not build when prices are low. They will make sensitivity (stochastic) analyses of their price forecasts. In particular, they make careful projections of the factors that can alter the future supply/demand balance because when capacity is expected to be tight, prices should be high, and when surplus capacity is expected, prices should be low.

Because the market model relies so heavily on price signals, the economic integrity of pricing mechanisms within the market rules is paramount. Prices should not only signal how much total capacity to build, but also how much of each type of capacity to build. Prices during peak hours should signal the profitability of peakers up to and no further than the point at which mid-load plants are a more efficient choice. Likewise for base-load plants. Prices should signal that new capacity is more valuable in some locations than others. The shape of the price duration curve should adjust, and the market should self-correct, if wrong decisions are made; for example, if too much base-load plant is built, then off-peak prices should drop. If there is too little capacity in a transmission-constrained region, then prices there should rise.

2. ***Reliability.*** In well functioning markets for other commodities, and ultimately in the new electricity industry model, the concept of reliability is also entirely different. As a commodity becomes scarce, prices rise and "low-value" customers low on the demand curve choose not to consume. This is illustrated in the simple supply and demand curve shown in **Figure 1.** When supply of the commodity falls, prices rise and more customers chose not to consume, as shown in **Figure 2.**

In well functioning markets for other commodities, prices always settle at a level where supply equals demand. There is never a situation

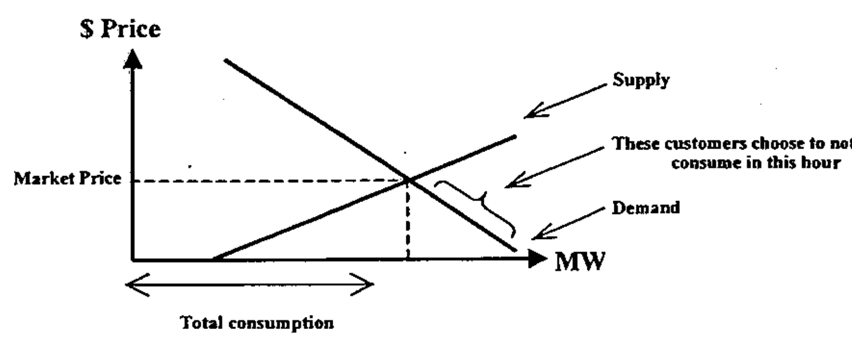


Figure 1: Simple Supply and Demand Chart, High Supply Hour

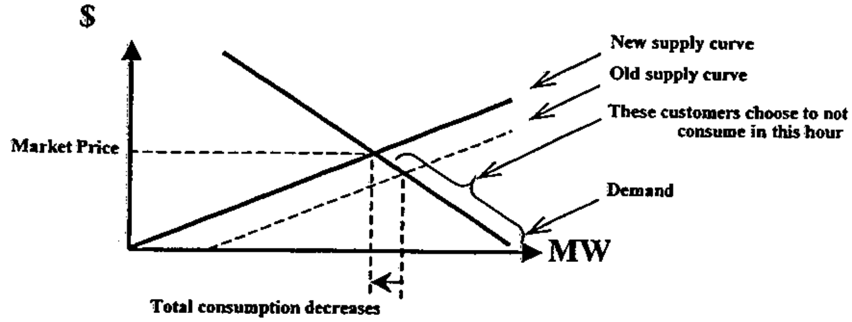


Figure 2: Simple Supply and Demand Chart, Low Supply Hour

in which a price is posted at a level that a customer is prepared to pay, but the customer is told that it cannot buy because of lack of supply; prices settle at the intersection of the supply and demand curves. Customers *choose* whether to consume or not at the market price. The choice is always theirs, not that of an independent system operator (ISO) or some other third party. Well-functioning markets are thus always "reliable" because the lights only go out if a customer wants them to, given the price.

Of course, in the new market model there is still a role for operating reserve and other ancillary services because of the uncertainty involved in real-time electricity production and consumption. Electricity travels at the speed of light, after all, and production and consumption must always be perfectly balanced.

C. In Theory: Consistency between the Old and New Industry Models

In theory, the new industry model should produce the same outcomes as the old. The new model begins with the concept that the market will be used to organize the production of electricity in a least-cost fashion and to establish prices based on that production. The market price signals will also be used to provide the correct incentives for long-term investment decisions.[2]

The theory is that if the market is organized so that all generating units bid their marginal cost of production to a system operator that dispatches plants and sets a market price based on the marginal bid accepted, then a least-cost dispatch will be achieved and efficient price signals will be provided for all production, consumption, and investment decisions. Generators will not have to guess at the clearing price in order to maximize profits and efficiency will not be distorted by incorrect guesses.

In such a system it is easy to see how the proper price signals to encourage short-run efficiency in production and consumption are provided. For instance, what bidding strategy would maximize short-run profits for a firm that was a pure price taker, i.e., had no ability to influence market prices? Under the new model such a firm maximizes profits by bidding at marginal cost.

- If the firm bids above marginal cost, there is a chance that the firm would not be dispatched and yet could have made profits at the market-clearing price.
- Conversely, if the firm bids below marginal cost, there is a possibility that the firm would be dispatched at a price that was less than its costs and would therefore make losses.

Marginal cost bids, with a single clearing price as described above, produce the proper price signals to encourage efficiency in production and consumption, produce incentives compatible with dispatch instructions, and maximize the short-run profits of a firm in a perfectly competitive market.

It is less easy to see how marginal cost bids and a clearing price based on these bids will provide the right incentives in a competitive market for development of an optimal amount of capacity, and capacity of the right type, as the old industry model did. The theory is, however, rather simple. **Appendix II** describes how decentralized market participants theoretically have incentives to build the same amount and type of capacity as central planners. The same basic economic principles apply to decentralized investment decisions as did to centralized ones—the difference is that the centralized model relies on internalizing quantity decisions and the decentralized one relies on external price signals, profit incentives, and the existence of well-structured markets.

D. In Practice: The Uneasy Relationship between the Old and New Industry Models

The problem is that what works in theory doesn't always work in practice. The motivation for moving from the vertical integration model stems, after all, from the practical problems and inefficiencies of central planning. Competition is the basis on which the US economy was built. But there are practical problems in the new market model for electricity, and particularly in the transition to the new market model.

The new industry model does not produce exactly the same outcomes as the old. There are several reasons why efficiency and reliability outcomes are affected.

1. Mostly, There Isn't an Active Demand Side. Those demand curves of Figures 1 and 2, in reality, are basically vertical. This means that the market cannot rely on demand response to produce market-based reliability solutions, and to lift market prices appropriately.

The demand side of the electricity industry has a unique characteristic that complicates restructuring and is often overlooked: Final end-use buyers of electricity are for the most part accustomed to paying tariff prices that are averaged, often over a year or more, yet the underlying cost of electricity varies tremendously from hour to hour. They have no incentive to turn the lights off themselves when supplies are scarce, because the price to them to keep on consuming is the same as when supplies are plentiful.

For markets to work properly there needs to be an active demand side in which consumers increase consumption when prices fall and decrease consumption when prices rise. Some electricity users have a very high value from every kWh they consume (heating for the elderly and sick, lighting for the streets, and so on) and others have a much lower value (pot 9 at the local aluminum smelter, for example). Yet all are consuming the same basic commodity transported for the most part on the same system. Customers and regulators demand high reliability for the high-value uses and in the old world this meant avoiding outages. Because of averaged tariffs, however, many of the larger, low-value users had little incentive to decrease consumption at times of shortage and high cost. The challenge in the restructured world is to keep the lights on for the high-value users, even when costs rise, and to give other users price signals that they are better off turning off their "lights" or consuming at different times of the day.

For markets to work, there needs to be an active demand side in which consumers react to price changes.

The way to meet this challenge, which is discussed in greater detail below, is to charge the low-value users on an hourly basis; and potentially, if more demand response is needed, to charge high-value users on an hourly basis. To do this, hourly meters are required, and hourly pricing is required.

Hourly charging doesn't mean we have to do away with price stability. Even with hourly pricing, price stability can still be assured through long-term contracts that average out costs to consumers over time. As long as these are structured appropriately—e.g., for fixed amounts and not on a full requirements basis—the incentives to respond to price changes will be the same as if there were no contracts. There will be just as much demand response potentially available, and hence market-based reliability.[3]

2. Mix of Regulated Capacity Requirements with Market Incentives in the New Model. The second reason for the uneasy relationship between the old and new industry models is more complicated: There is a messy mix of regulatory and market incentives.

Hogan described this problem well when he said that the development of rules for reliability and rules for markets confront a familiar human condition: the desire for mutually exclusive things.[4] Those involved in industry restructuring in the United States want to be able to allow multiple control areas with different approaches; to maximize opportunities and flexibility for choice, trading, and commerce; to mini-

  1040-6190/01/$–see front matter PII S1040-6190(01)00249-4

mize commercial interference; to maximize economic efficiency; and to improve—not worsen—reliability. But what we have is an interconnected grid where everything affects everything else, fast. We have conflicting agendas and ideologies, too many regulators, and too many good ideas. We have socialized reliability solutions: NERC, regional and state reliability councils, and even the ISOs/RTOs themselves, are still running on their old agendas, continuing to apply old-structure reliability and installed margin rules on top of the new market institutions and on top of the market-based reliability solution. Furthermore, all sorts of uncoordinated players, agencies, regulators, and others also have their own intellectual frameworks for the transition. Each state, or at best group of states, has more or less gone its own way. The result has been a hodgepodge; the old industry model rules are at times arcane within the new system.

The basic problem is best illustrated by going back to look at how reliability is planned for in the old model, and comparing that to the new. In theory, the old model works by engineers trading off an assumed value of lost load (VOLL) and expected loss of load probabilities (LOLP) against the cost of increasing installed reserve, in order to target an "optimal" level of reserve. They then build those reserves so that the target level of reliability is attained, as indicated in **Figure 3**.[5]

The new industry model works in a similar way, in theory, except that everything is decentralized. Each customer declares a maximum reservation price it is prepared to pay (the demand curve is just an ordered list of individual customer VOLLs), and, as Figure 2 shows, prices must rise in order for customers to choose not to consume. The occurrence of occasional high prices is built into generators' expectations, and ultimately into the quantity of capacity they decide to build. The quantity of capacity built is thus consistent with customer value of lost load.

The new industry model relies on customers individually specifying their reservation prices (VOLLs), and this is where it comes unstuck. For reasons described in the following section, very few customers are doing this. Even more importantly, few systems are in place to allow price signals, or reactions to price signals, to get through. Consequently, two mechanisms end up trying to achieve the same result and neither works properly:

- The old model treats reliability as a common good. The responsibility for providing it is socialized: Everybody shares the cost of providing the reserves that central planners deem to be necessary.
- The new market-based model is based on individuals taking personal responsibility, and hence the conflict: Consumers are reluctant to take on responsibility themselves when they can be assured that "big brother" is looking out for them.

On the one hand, policymakers want a proper market, with consumers taking personal responsibility. On the other hand, they are discouraging them from doing exactly that. The installed reserve margins, which exist in the Northeast markets, for example, act like "oil on water,"[6] smoothing out energy prices. Even if there were plenty of hourly meters and customers with hourly prices, the price signals would be dampened.

There are other practical implementation problems with the over-

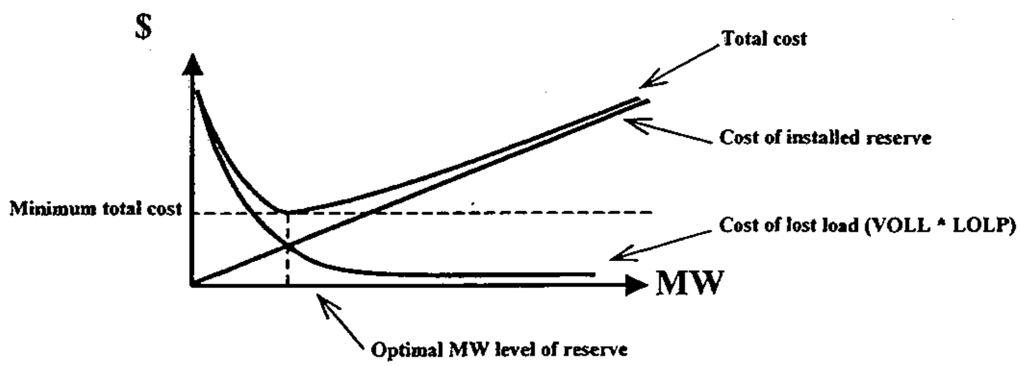


Figure 3: Optimal Level of Reserves, Old Industry Model

lap of these two models. Installed reserve margins are one thing to manage if the same company that sells electricity to the final consumer is vertically integrated and must also generate the energy. But they are altogether different in a world where electricity retailers can be asset-less marketers whose peak load is determined by the success or otherwise of an advertisement at half-time in the Super Bowl. All sorts of new incentives, rules, measurements, and policing mechanisms are needed. Pennsylvania-New Jersey-Maryland (PJM) is the most advanced market is this regard, and it is still working through all the issues. Its capacity market is yet to be truly tested.

Geographic problems also exist. Where a company in an old industry structure RTO (regional transmission organization) trades with one in a fully market-based RTO, both sides can have distorted incentives. Two transitioning RTOs side-by-side can have distorted trading incentives as well. Capacity rules that smooth energy prices in one system can, for example, have the effect of increasing demand during peak periods for exports to neighboring systems that don't have the same capacity rules. PJM, for example, has had problems of importing energy price spikes from neighboring systems. Capacity rules in systems like PJM insist on minimum levels of installed reserves, and the revenues from providing installed reserves give a price signal to potential builders of new capacity. An alternate method of addressing long-term reliability is to rely solely on price spikes in the energy market to provide incentives to supply on high-demand days, to provide price signals to reduce consumption, and to provide price signals to potential entrants. But this can be a problem for PJM because it is difficult if not impossible to separate its reliability solution from those of its neighbors, who tend to rely to a greater extent on price spikes and who can take advantage of importing energy from PJM's capacity in peak hours.

Going cold turkey and relying prematurely on the new model can result in California-style reliability nightmares.

This decreases the effectiveness of the PJM reserve margin, leading to a requirement for increased reserves, and so on.

In short, a mixture of old-model incentives and new-model ones can stall the transition to the new market-based model, yet, unfortunately, going cold turkey and relying prematurely on the new model can result in California-style reliability nightmares.[7]

3. Other Methods Are Needed to Raise Prices. The third reason for the uneasy relationship between the old and new industry models is that without an active demand side, marginal cost-based prices are not high enough.[8] As described earlier, with competition among generating units, generators in a pool have an incentive to offer energy at their variable cost. In order to make sure that generators are economically viable, however, it is necessary that prices are sometimes higher than the variable cost of the last generator taken.

For example, on the peak hour of the year the price of electricity must be higher than the variable cost of the last peaker running, or it might never have an opportunity to recover its fixed costs, and that peak generator is required to run in order to maintain reliability.[9]

Appendix III describes this in the context of the "cross-over" charts of Appendices I and II. The practical result is that some mechanism must exist in the market model to sometimes lift generator remuneration above price levels based on marginal cost.

Demand response would fix this, if it existed, because prices rise in times of shortage until enough people choose not to consume. *Demand response behaves exactly like a "peaker without a fixed cost."*[10] An active demand side means that when the amount of capacity offered is less than the amount of capacity required to securely serve load in a dispatch interval, the price will rise in order to elicit a reduction in demand. The price level required to achieve the necessary reduction sets the clearing price. In this way, the price rises above the marginal cost of the most expensive marginal unit, as indicated in **Figure 4.**

  1040-6190/01/$–see front matter PII S1040-6190(01)00249-4

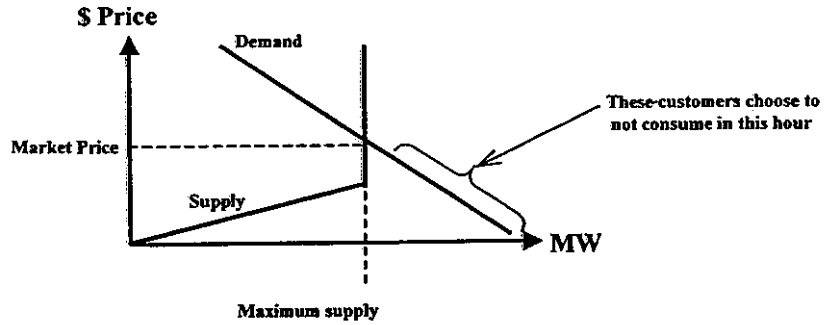


Figure 4: Prices Rise and Consumers Respond to Supply Shortage

When the price rises frequently enough or high enough that it is profitable to construct more capacity, the supply side of the market should respond. Since the price increase in these periods when all supply is scheduled exactly reflects the value of electricity used by customers at the margin, the exactly right amount of capacity development should be encouraged by the pricing mechanism.[11]

Of course, as was described earlier, an active demand side means that sometimes, even when there is more capacity available, price can be high enough that some customers choose not to consume. This outcome, as shown in **Figure 5**, is just a normal interaction of supply and demand, familiar in other industries. The resultant prices are still those that are exactly consistent between consumer willingness to pay and producer willingness to sell.

But if demand response is inadequate, as it is right now, an alternative solution is needed to raise prices above marginal cost. Available alternatives are discussed later.

III. Demand Response

The root of all the problems just described is demand response. Reliability and efficiency problems will not go away until electricity markets can be counted on to provide reliability and long-term investment solutions by themselves, without the need for outside interference. Other markets in other products do this fine, and it is time electricity markets did also. Other markets do this because they have active demand sides as well as supply sides.

The fundamental problems preventing an active demand side from happening are twofold. First, there are technical problems. In order to charge hourly prices, hourly meters are required at an end-use level. Hourly meters are not yet widespread and, furthermore, even if they were, consumers might not be able to respond quickly enough to changes in price. Second, there is considerable inertia to overcome. Most electricity consumers, even some of the larger ones, are simply not used to modifying their consumption patterns based on price.

Until they are, central planners and system operators still need to count on socialized reliability solutions to avoid turning customers' lights off without their permission.

A. What Is the Problem with the Demand Side Today?

In order for demand responses to efficiently set price during periods when all capacity is dispatched (or scheduled as reserve) it is necessary that end users face the actual price and are able to respond to the price. This is a difficult proposition. In order for the demand to respond optimally to price during these periods, a large amount of load has to actually *pay* the spot price for electricity use during these periods or be able to *save* at the spot price rate for reducing demand during these hours.[12] This requires end user metering and pricing that is not common. It requires a fundamental change in the way people think about pur-

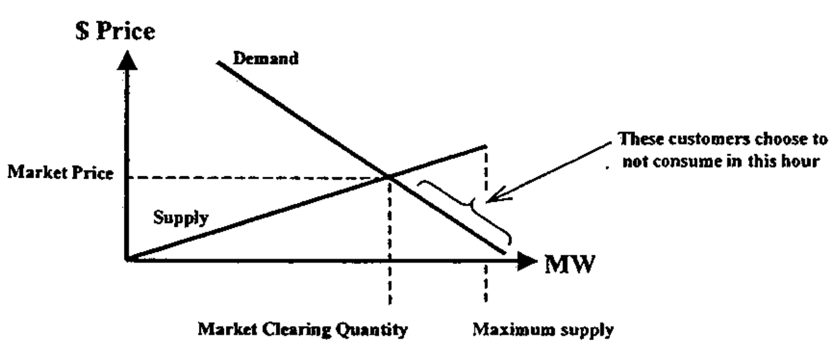


Figure 5: Normal Supply and Demand

PII S1040-6190(01)00249-4

chasing electricity.[13] It also requires a willingness on the part of regulators and politicians to allow many consumers to take responsibility for their own purchasing decisions. Regulators and politicians tend to take a paternalistic attitude regarding shielding consumers from price fluctuations.

Further, such programs cannot be implemented in an effective manner quickly. In order to be able to respond, advance planning and technology is needed. Even if customers collectively have a stunningly altruistic bent (California Governor Davis' gesture of turning off the Christmas tree lights outside the California capital buildings in 2000 comes to mind) the infrastructure is currently insufficient for large numbers of customers to respond quickly to specific situations of scarcity. The problem is that electricity prices can vary substantially even within a day, and often at very short notice. Day-ahead markets help, especially for industrial customers, because they help customers plan consumption ahead of time. Customers can specify a maximum (reservation) price they are prepared to pay, and can plan to consume if the day-ahead market price is lower than this price. They are protected against spot price movements. The deepening of forward markets in general, including longer-term bilateral markets, is part of the solution to improving demand-side response. The spot market will always remain important, however, because all forward sales are priced based on the alternative of not buying or selling in the spot market. Ultimately, if end-user hourly metering and pricing is not in place for a substantial amount of spot market load, and if end users have not anticipated potentially high prices and invested in technology and programs that will enable a response, demand will remain very inelastic.

The designers of the wholesale market in California planned on demand response, but AB 1890—a legislative enactment—froze retail prices. **Appendix IV** describes how a lack of demand response was a critical reason, perhaps the most critical reason, for the California disaster.

A lack of demand response was a critical reason, perhaps the most critical reason, for the California disaster.

B. How Should the Demand Side Participate?

That buyers should pay the market-clearing price—or a contract price based on it—is a condition of all competitive markets in other commodities. Assuming this to be possible with electricity (putting aside the California experience), how can it be achieved?

1. How Much Hourly Metering Is Needed? Some electricity markets overseas have actually required that all consumers who choose retail access must install real-time meters. These meters track a consumer's electricity use hourly (or more frequently) and some of them can provide data to the customer in real time to facilitate the customer's response to hourly market price variations. Such meters typically include telecommunications equipment to send the hourly usage information to the system/market operator as well.

The alternative to hourly meters is load profiling. Load profiling can be used to estimate how much energy each consumer uses in each hour. Use of load profiling to determine the hourly usage by a particular customer involves developing typical load profiles that are assumed to represent the patterns of electricity use over the month[14] of each customer to whom the profile is applied. In effect, the profile is scaled to match the total monthly consumption of a particular customer, and then assumed hourly loads are multiplied by the hourly spot prices to establish the supply portion of the customer's bill. The actual procedure is normally to weight the hourly spot prices by the load profile, compute a weighted average price, and charge that price for all consumption by a customer within the class to which the load profile applies.[15]

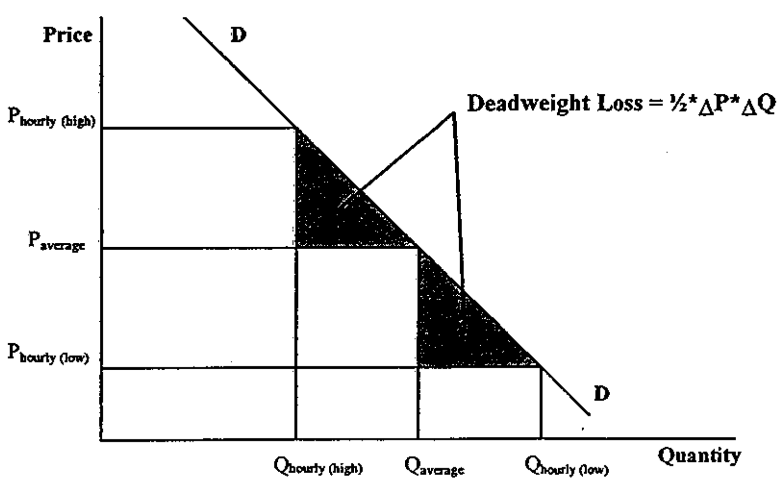
With load profiling the customer does not directly see the price signals in the market. A load-profiled customer will pay a price that depends upon both the load profile of his group (which may or may not be known in advance) and upon the spot prices in each hour of the billing cycle. Because the customer's bill is not directly related to his own consumption in the high-cost and low-cost hours, the efficient signal of hourly spot prices is very substantially diluted.[16]

What is the right mix of hourly metering and load profiling? If there were no cost differences between using hourly meters and using load profiles, it would be efficient to put all retail access customers on hourly meters. Hourly meters are more expensive, however, so there is a tradeoff. Some customers may be better off on hourly meters, while others might not. The tradeoff is between the extra cost of the hourly metering, and the extra benefits of hourly metering. The benefits come from two places:

1. Benefits from reducing the societal losses from charging too high a price in some hours and too low a price in others; and
2. Benefits from reducing the cost of socialized reliability solutions.

The first of these benefits can be determined using standard cost analysis,[17] as illustrated in **Figure 6**. Given a customer's demand curves for electricity, it is possible to determine the societal losses. These losses are the shaded triangles labeled deadweight loss. If the deadweight losses exceed the incremental cost of hourly metering, the customer should be required to have an hourly meter.

The second benefit of hourly meters is a social one, and increases the number of hourly meters that should be installed beyond those that are already justified on a stand-alone basis from the first benefit. For every hourly meter installed (with hourly pricing) to a customer that is actually price-responsive, the requirement for expensive socialized reliability solutions like an installed capacity margin falls. (Or in the case of California, which has no capacity requirement, the extent of price spikes and the likelihood of blackout falls.) Eventually things like installed capacity requirements can be eliminated entirely.[18]



Note: The analysis represented here assumes ΔQ is relatively small and the supply curve is relatively flat (compared to the demand curve) over this range.

Figure 6: Cost Analysis of Profiling

In general, of course, these benefits are difficult to quantify. But while it would be useful to have enough information to make detailed assessments, it is not necessary to have detailed information on customer demand characteristics to begin to define classes for whom metering might be cost-effective: it is likely to be the case if (1) the customer is a large electricity user, (2) the customer's price responsiveness (elasticity of demand) is high, or (3) the difference between the hourly prices and the load profile price is large. The costs of hourly metering? Recent estimates have been that the annualized cost retrofitting a high volume of standard meters to store hourly information is as little as $12 per meter. If the hourly metering is not cost-effective, load profiling should be used, combined with interruptions when spot prices are high.

2. ***What Does It Mean to Be "Price Responsive"? Do Customers on Hourly Pricing Need to Bid into the Market? Do These Customers Need to Know the Price in Real-Time?*** If some buyers indeed pay hourly prices and have their hourly consumption metered, it is desirable that they actually make bids to the market specifying the quantities they are prepared to buy and the reservation prices they are prepared to pay. Furthermore, it is desirable that the market is capable of processing this information and accepting or rejecting bids, as appropriate depending on market price, in a timely manner. But it is

PII S1040-6190(01)00249-4

not necessary that all buyers make bids. Others can respond to price signals by simply not consuming.

The ease of achieving large numbers of consumers that bid depends on the time frame in which electricity is purchased. When it is purchased in forward/day-ahead markets, it is relatively straightforward. In a pool, the market operator can evaluate the bids against the day-ahead generator offers and can produce a set of prices and scheduled production and consumption quantities consistent with all the bids and offers. Depending on whether its bids are accepted or not, a buyer can plan its consumption for the following day.

For energy purchased in real-time (in spot markets), it is not so straightforward. Energy is always traded in spot markets because forecasts are never perfect. Spot prices are much more volatile than forward prices because things can change at short notice. And spot prices are typically calculated after the fact, once actual system conditions are known. Therefore, in order for a system operator to be able to curtail a bid in real time based on a reservation price (i.e., "dispatch load"), sophisticated software and real-time control equipment is necessary.

It is beneficial if large MW quantities of load are price-responsive and under real-time control of the system operator because it adds to the short-term security of the system; as reserves fall, prices rise and some bids are consequently not accepted. Yet it is equally valuable from a system security point of view if customers find alternative means to respond to rising prices. It is not necessary to rely on all price-responsive load being dispatchable and bid into spot markets. If buyers even *anticipate* high prices and have the ability to respond by not consuming, then they will also contribute to a market-based solution to the longer-term reliability issue. It *is* necessary, however, that sufficient buyers pay the hourly prices for the energy

they do consume. Otherwise they won't ever face hourly spot prices for the imbalance between their forward purchases and their actual load, and they won't have incentives to respond to changing prices.

There are numerous mechanisms for anticipating prices. These mechanisms differ by how far they are in advance of the dispatch hour:

- At one extreme, experience of market prices over the long term indicates to buyers that electricity tends to be more expensive in summer, in the afternoons, and especially on hot days.[19]
- Day-ahead markets, where they exist, allow buyers to plan for their consumption in advance of the real time, and to lock in purchase prices at day-ahead prices, which are a forward projection of real-time prices.
- Some markets produce rolling prices between the day-ahead and the dispatch hour. Depending on the specific market, these can either be advisory prices or updated prices under which energy trades. Either way, they can signal the value of demand response right up to the dispatch hour.
- Calculating and communicating the price just prior to real time or having dispatchable load in real time is the final means of stimulating a demand response, and the only one that requires knowing the price in real time.

C. Socialized Reliability Solutions

In the absence of an active demand side, or in the interim until one develops, what are the alternatives? Since the mechanisms for price response by consumers do not yet exist properly as they do in other markets, we are forced into "fix-it" solutions to ensure that reliability is maintained (where "reliability" is defined in the old model terms of socialized responsibility and means ensuring that the lights stay on for everyone some predetermined percentage of the time).

Numerous solutions to this exist in the United States and around the world. They are summarized below.

1. LOLP/VOLL Methods. LOLP/VOLL methods have at various times been applied to overseas markets, most notably the

  PII S1040-6190(01)00249-4

original incarnation of the England and Wales pool, which used such a method for many years, and also in a modified form in the market in Argentina. The LOLP/VOLL method is a solution that can be applied when there are forward energy markets run by a centralized market operator.

LOLP/VOLL methods work by incorporating in the energy price an additional payment to reflect the extra value of capacity at times when capacity margins are low. By multiplying the loss-of-load probability, or LOLP, by the value of lost load, or VOLL, and adding it on a weighted basis to forward energy prices, the method approximates the expected value of an outcome where prices are occasionally set by demand-side response. It is used in situations where demand is inelastic; the demand-side response is represented by the value to an "average" consumer of not being subjected to an involuntary curtailment. The LOLP is a forward estimate of the likelihood of curtailment.

Both LOLP and VOLL are estimated parameters. Although there is always a finite probability of loss of load, for all practical purposes LOLP is zero when supplies are more than 10 percent above the expected demand. Under the theory of LOLP/VOLL pricing, in spot markets there is either loss of load or there is not, so the value of LOLP is either zero or one. Forward market prices and spot market prices are thus consistent on an expected value basis.

VOLL is a concept widely used in the electricity industry. The correct value to use is the average value to all customers, if disconnections will be random, or the marginal value to those who will be cut off, if disconnections will be specified. Its value can be estimated, from the value-added in each industry using electricity, and also from other methods.

The LOLP/VOLL method provides a reliability solution by increasing prices when demand is

close to available generation capacity. This provides additional compensation to generating units that are available during periods when supply is tight. Generators are thus encouraged to invest in additional generation capacity, in the expectation of earning higher prices in the market through the mechanism than they would if the price-adder didn't exist.

This mechanism also has the benefit of encouraging generating plants to be available when they are most valuable, i.e., when the likelihood of shortage is greatest, and of encouraging large consumers, who pay the price hour by hour, to reduce consumption in those hours. The adder is also paid to providers of reserve, who might not be generating in an hour, to be available when the value of their reserves is highest.

VOLL values are commonly estimated between $1 and $5 per kWh. Although there is a wide range of estimates of VOLL, the efficiency of the system is not particularly sensitive to the actual number chosen, so long as it is fairly high. The payments to the generators and the cost to customers is the sum of all LOLP*VOLL payments, and if there is slightly more capacity there will be fewer episodes of high LOLP.

2. Installed Capacity Obligations. Installed capacity obligations and accompanying markets for installed capacity are a method of reliability solution used in the Northeast United States. Load-serving entities (LSEs) are required to own sufficient "capacity tickets" to match their peak load, plus an installed reserve margin designed to provide a predefined level of system security. Tickets can be obtained by owning physical capacity, or by buying tickets from other owners in annual, monthly, or perhaps daily capacity markets. Generators for which tickets have been issued must meet certain conditions. They must be demonstrably operable and pass certain operability tests, they must agree to a means of centralized coordination of planned outages, and they must agree to curtail out-of-system sales under certain conditions such as emergencies. LSE peak load may change from time to time in an environment of retail access and the holders of tickets may change, so the ISO keeps track of the net position of each LSE. Deficiency penalties are applied to LSEs that don't have enough tickets.

Installed capacity margins are designed to ensure reliability by specifying a required *quantity* of reserves, which the capacity ticket market converts into a price for capacity. (The required quantity of reserve relates to the required level of reliability.) This method works in the opposite way to the LOLP/VOLL method, which specifies a *price* for reserves from which the market induces the quantity of capacity to be built. The installed capacity model tends to determine capacity prices so as to remunerate the fixed costs of keeping units available (such as peaking units) that are not on average otherwise expected to profitably operate in a given year.

Installed capacity markets are not perfect and suffer some problems: How do you deal with customers switching supplier during the year? How do you know the peak load of an LSE before the fact? (Capacity tickets must be arranged before the fact, but actual load is not known until after the fact—there are incentives to under-forecast). How do you efficiently command energy production from capacity in other systems with other system operators and transmission limitations at short notice? How do you discount the value of capacity in other systems? The methodology suffers from policing problems: How do you know capacity is really there when it is needed? How do you make sure capacity is not sold or contracted to another pool (especially if the capacity is located outside the pool in the first place)? What is an appropriate noncompliance penalty? What are the incentives to ignore export recall instructions? Yet despite their imperfections, these issues are being worked through and installed capacity markets do show signs of achieving their goals.

A variant on the installed capacity model is to apply very high

operating reserves in place of installed reserves. This variant has some advantages in avoiding the problems of installed reserve obligations but has efficiency problems in that it requires large amounts of capacity to be continuously available but not called upon.

3. Other Methods. Some overseas competitive market designs have responded to the capacity problem by introducing capacity payments in the form of a specific additional kWh payment made to generators who are called to run at pre-specified times when capacity is expected to be tight. This system is similar to the LOLP/VOLL method, but more arbitrary in the application of the price adder. Unfortunately, this system has the effect of inducing generators to reduce their bids so that they run and get the payment, and thus largely fails to increase prices to induce extra capacity. An alternative is a payment of an annual payment per kW installed, but this has the disadvantage of not sending the right signals to generators to be available when capacity is actually tight.

Another alternative is a legislated solution requiring LSEs to own installed capacity to meet their load, and own or contract for capacity to meet their share of reserve requirements. This solution was implemented by Greece in 2001.[20] It follows the same basic idea of the installed capacity market, but because of the ownership requirement eliminates some of the problems of capacity tickets that stem from them being traded. This solution requires continued vertical integration, so it resembles the old vertical integration model in the United States. The ownership requirement makes entry to the retail market more difficult.

Reliability solutions for specific circumstances have also been suggested overseas. "Dry-year price-insurance" was suggested in Mexico[21] to smooth payments between wet and dry hydro years, and a "K factor" payment was mooted there to account for perceived investment risk premiums required in the transition to a fully restructured market.

4. Do Nothing. The final alternative is to do nothing; to rely only on generators bidding up prices above marginal cost in times of scarcity, although not to the point where it constitutes market power. This is essentially the model that was effectively implemented in California after AB 1890 imposed a rate freeze. It is also essentially the model deliberately implemented in New Zealand—but on an island where there is nowhere else for the energy to go, and in a system with substantial hydro facilities, there is much less of an issue. Even so, New Zealand recently had hydro-related shortages. But the experience in California is a warning against doing nothing.

IV. Conclusions

The goal is ultimately to avoid socialized reliability solutions.

Socialized reliability solutions such as installed reserve margins are not required once there is an adequate demand-side response. They must be regarded as transitional, smoothing the way from the old model to the new. Ultimately the goal has to be to have personal responsibility—i.e., market-based reliability—and efficient investment signals provided by an active demand side. With this, workable reliability will be assured so long as enough customers (1) are metered and charged on a hourly basis, (2) will defer or stop consuming when prices reach a reasonably high level, and (3) have the technical ability to indeed stop consuming when prices reach that level.

The need for socialized reliability solutions in the interim comes about as a result of the difficult transition from the old industry model to the new and in particular the lack of an active demand side. The fact that there has not been a consistent answer to the question of a socialized reliability solution alludes to the complexity of the problem. The role of capacity payments and the appropriate form of the solution are among the most contentious issues in the field of wholesale market design. Socialized reliability solutions all have difficult issues, but if we can't get meaningful demand response quickly, and experience is telling us we can't, there is no avoiding them.

In the longer term, an active demand side is critical to a successfully restructured electricity industry. It means that the industry behaves like other markets; reliability and efficiency are assured, not things that must be planned for. The designers of the California wholesale market knew this when they created the market—they wanted an active demand side and so didn't build LOLP or capacity markets or other things like other places—but AB 1890 pulled the rug out from underneath them. An active demand side does not happen by accident, but disastrous reliability consequences can. ■

Appendix I: Generation Planning in the Vertically Integrated World[22]

The analysis used to determine which types of plants to build in the vertical integration model can be represented by the following illustration of a load duration curve (**Figure 7**). If load is represented over a year by this curve (sorted by high-load hours to low-load hours) then some plants will run to meet load almost constantly (C = base-load), others will run about half the time (B = mid-load) and others will run only in the highest load hours (A = peakers).

It follows that it may be cheapest to have a combination of generating technologies combining to meet the total load requirements. Base-load plants like large coal-fired units tend to have high fixed costs, but low running costs, so they can have the cheapest average cost when they run for much of the year. Peaking plants like gas turbine (GT) or jet units tend to have low fixed costs, but high running costs; they are the cheapest option if the plant only runs for a few hours, but their average cost would be too high if the plant was run all year. Mid-load plants, like combined cycle gas tur-

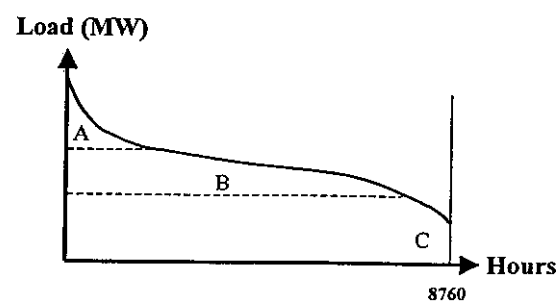


Figure 7: Load Duration Curve

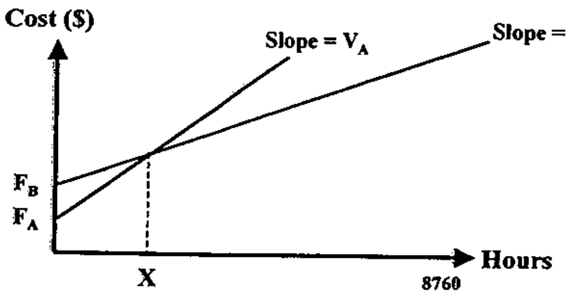


Figure 8: Run-Time of Last Peaker Built

bines (CCGTs), are somewhere in the middle.

To solve the problem of exactly how much of each type of plant to build, central planners use complicated optimization software. Conceptually, though, the solution is simple to find (at least while other factors are held constant). Assume, for example, that the three competing technologies have fixed costs of F_A, F_B, and F_C, and variable costs of V_A, V_B, and V_C dollars per MWh, respectively,[23] and that you can build these plants in any small increments you want.

Step 1. Start with the decision of how much peaking capacity to build.

One plant has to be built to run only in the peak hour of the year. (Since every other hour has less load, it won't be needed in those hours.) Assume that the cheapest plant to run this hour is indeed the peaker; i.e., $F_A + V_A < F_B + V_B$ and $F_A + V_A < F_C + V_C$. In this case, the answer to the question, "How many MW of the peaker do I build?" is derived from a straight-forward cost analysis: peakers are the least-cost option so long as $F_A + XV_A < F_B + XV_B$, and solving for X gives us the number of hours that the last peaker built runs. Visually, this can be represented as a simple "crossover" chart, as in **Figure 8**. If a generator has to run for less than X hours, then the cheapest alternative is a peaker. If it must run for more than X hours, then a mid-load plant is cheaper.

It remains, therefore, to determine how much peaking capacity is required. If the marginal peaker runs for X hours, we can extrapolate to the load-duration curve to determine how much load must be met by generators that run for less than or equal to X hours, and thus the necessary MW of peaker capacity, as in **Figure 9**.

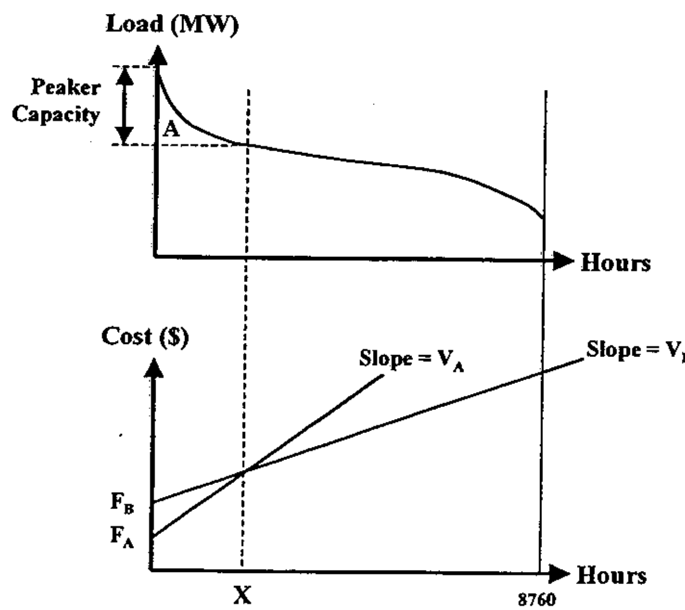


Figure 9: Calculation of Required Peaking Capacity

Step 2. The next question is how much mid-load capacity should be built.

In this case the answer is derived from a cost analysis of mid-load plants vs. base-load ones. Mid-load plants are the least-cost option so long as $F_B + YV_B < F_C + YV_C$, where Y is the number of hours that the last mid-load plant built runs. Visually, the crossover chart takes the shape of **Figure 10**. Following the same process of look-up to the load duration curve, we can see how many MW of mid-load capacity are necessary.

Step 3. The final question is how much base-load capacity to build.

Since there are only three types of plants in this illustration, it follows that any capacity not built as peakers or mid-load be built as base-load. The required and least-cost MW of base-load capacity to build under this old model method of calculation therefore equals

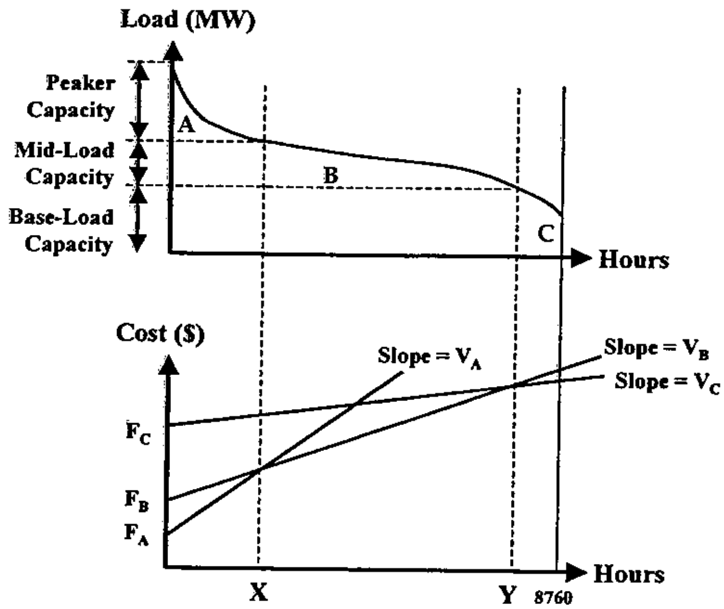


Figure 10: Calculation of Required Mid-Load Capacity

peak load minus peaking capacity, minus mid-load capacity.[24]

Appendix II: Market Signals Give the Right Incentives

It can be illustrated that marginal cost bids and a clearing price based on these bids will provide the right incentives in a competitive market for development of an optimal amount of capacity, and capacity of the right type, by returning to the central-planning crossover charts. In the illustrations that follow, a simplifying and artificial assumption is made that the fixed cost of the peaking plant is zero. (This is an important assumption upon which the results rely and the significance of it is returned to later.)

Step 1. Start with the evaluation of how much peaking capacity will be built.

If the peaker has a fixed cost of zero, and generator bids and market prices are based on marginal cost, then the peaker that runs only in the peak hour of the year will make an economic profit of zero: $(P - V_A) = (V_A - V_A) = 0$, where P designates the market price. (The term "economic profit" refers to profits in addition to those necessary to remain in business—an economic profit of zero doesn't mean that the generator will go bankrupt.)

The answer to the question, "How many peakers will be built?" is determined by a profitability analysis of mid-load plants: Mid-load plants will be enticed into the market and will undercut peakers at the level of output, X, where they can first recover all their costs, fixed and variable. This occurs where $PX = F_B + XV_B$, or put another way, the crossover point between peakers and mid-load plants is defined by the

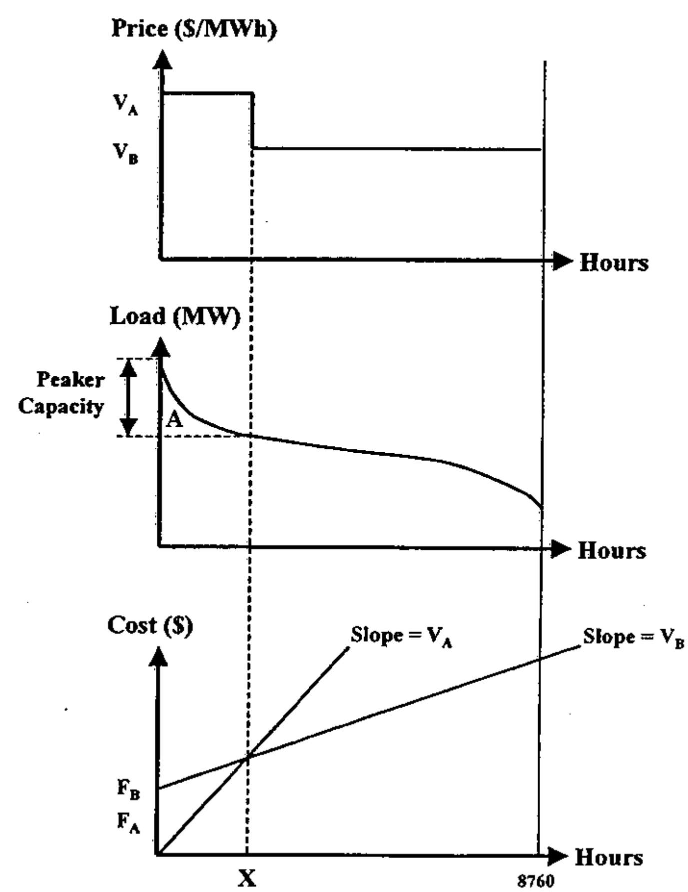


Figure 11: Evaluation of How Much Peaking Capacity Will Be Built

function $XV_A = F_B + XV_B$, as shown in **Figure 11.**

What economic profits do mid-load plants make when the enter? The first MW of mid-load plant type built makes just enough margin above its variable cost in the peak X hours to exactly offset its fixed cost since $PX = F_B + XV_B$. Every other MW of mid-load plant that is built and runs recovers its fixed costs from running in the same hours. In each additional hour that these units run beyond X, they are price-setters and thus receive a price equal to their marginal cost. Economic profits therefore remain at zero. Since no peaker runs beyond X hours, the economic profit of peakers stays at zero.

Step 2. Evaluating how much mid-load capacity will be built.

The answer to the question, "How many mid-load plants will be built?" is determined by a profitability analysis of base-load plants: Base-load plants will be enticed into the market and will undercut mid-load plants at the level of output, Y, where they can recover all *their* costs, fixed and variable—i.e., when the revenue from running in Y hours first equals or exceeds $F_C + YV_C$. Since the revenue from running in Y hours equals $F_B + YV_B$ (because mid-load plants make zero economic profits), the crossover point between mid-load plants and base-load plants is defined by the point where $F_B + YV_B = F_C + YV_C$, as in **Figure 12.**

What economic profits do base-load plants that enter make? The first MW of base-load plant built makes just enough margin above its variable cost in the first Y hours to exactly offset its fixed cost since the revenue from running in Y hours equals $F_C + YV_C$. Every other MW of base-load plant that is built and runs recovers its fixed costs from running in the same hours. In each additional hour that these units run beyond Y, they are price-setters and thus receive a price equal to their marginal cost. Economic profits therefore remain at zero, and hence the economic profit of all generators built is zero.

Step 3. Evaluating how much base-load capacity will be built.

As in the old, centralized model, base-load plants will meet all the remaining capacity requirements not met by peakers and mid-load plants.

X and Y, the running hours of the marginal peakers and mid-load plants built, are arrived at in an equivalent manner to the centralized planning model, except this time through decentralized price discovery and profitability analyses. Through these processes, market participants have incentives to build the same amount and type of capacity as the central planners did.

And generally speaking, if for any reason the market does not produce the right allocation of capacity, market participants will have profit incentives to self-correct the allocation. For example,

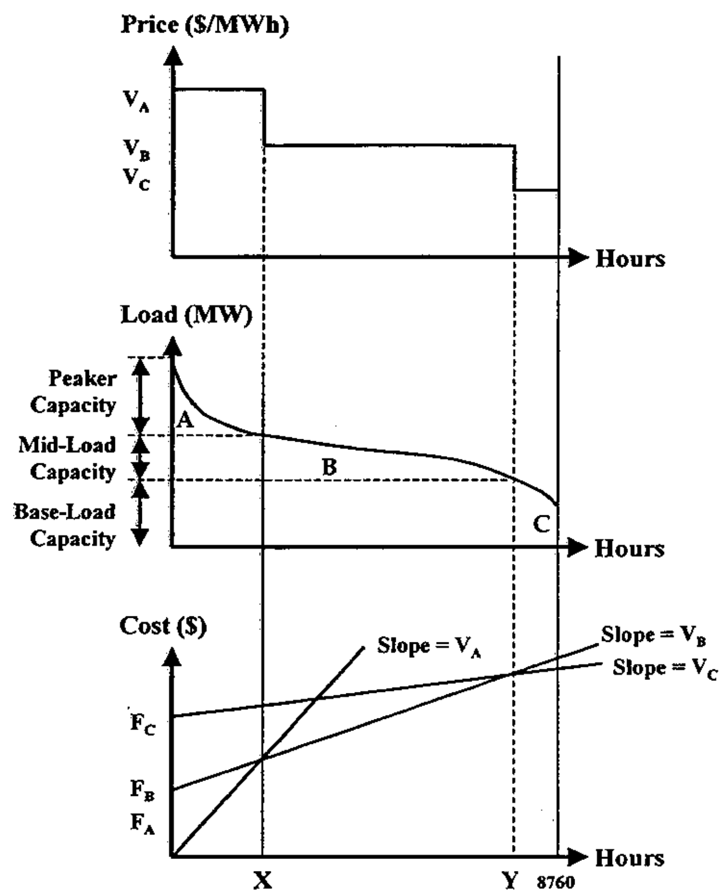


Figure 12: Evaluation of How Much Mid-Load Capacity Will Be Built

  PII S1040-6190(01)00249-4

if the proportion of peakers is too high, then existing mid-load plants will have economic profits greater than zero and other new mid-load plants will have incentives to enter. This would push down the prices to mid-load plants until their economic profits are zero, and no more enter. The opposite would be true if there are too many mid-load plants and too few peakers.

In an analogous way, market participants will have the correct incentives as to where new generators should be located.

Appendix III: Demand Response Is the Peaker without a Fixed Cost

In the comparison between the old and new industry models in Appendix II, the opening assumption was that the peaker has no fixed cost. What if it does?

Mid-load plants are encouraged into the market when $PX > F_B + XV_B$. But if P is too low, by not taking into account F_A, then fewer mid-load plants will be viable and fewer mid-load plants will thus be built. This can be seen by transforming Figure 9 as in **Figure 13**.

Central planning tells us that the efficient entry of mid-level plant occurs at X hours, yet if the price in those X hours is set only by V_A, and not also F_A, then decentralized incentives will not induce mid-load plants until they can run for at least X′ hours. Some of the mid-load plants that in reality are the best choice for the system will not be built because of the low price signals.

The problem works its way down the list of plant types. The end result is an inefficient capacity mix. Of course, the size of the inefficiency depends on the size of the fixed costs concerned. Demand response solves the problem by behaving like a "peaker without a fixed cost." The slope of this demand "peaker" is the price (VOLL) at which consumers are willing to stop consuming.[25] The most efficient number of outage hours is thus a function of VOLL and the fixed and variable costs of a physical peaker. Consequently, when prices are allowed to increase to VOLL during an outage, prices exactly recover the fixed and variable costs of the mix of peakers, mid-load, and base-load plants built, and load is unserved in hours when the cost of serving it would exceed VOLL.

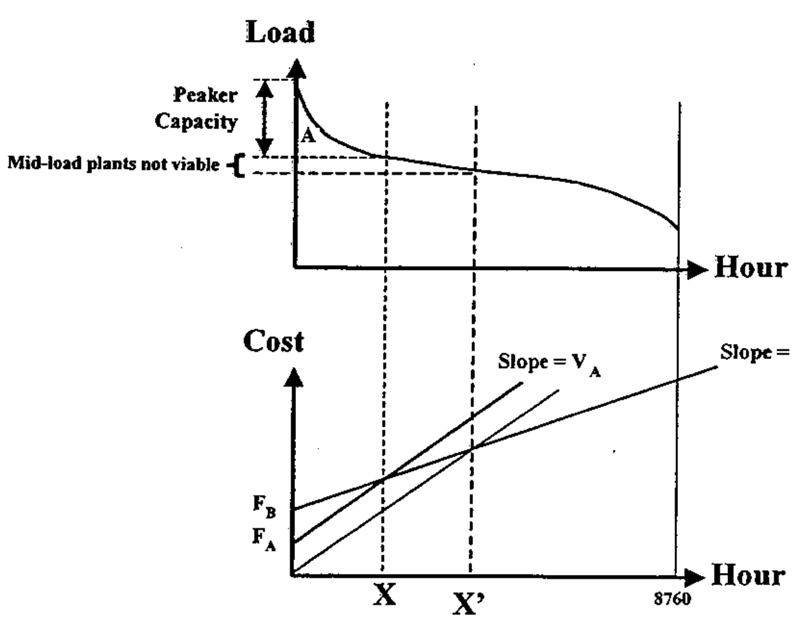


Figure 13: Entry of Mid-Load Plants If Price Is Too Low

Appendix IV: Case Study—The Experience in California

The lack of an active demand side is a principle contributor to the problems in California.[26] Following the publication of the Blue Book in 1994, efforts began to establish a wholesale market that would facilitate restructuring in California. The California Public Utilities Commission (CPUC) provided direction as to how that market would work in D.95-12-063 and D.96-01-009 and set forth a market model that envisioned dispatch locational marginal costs defining locational prices and envisioned active demand-side bidding. Demand-side bidding was the vehicle that would be used to provide capacity cost compensation to marginal generators as opposed to the installed capacity requirements in other U.S. markets in the East and as opposed to the LOLP/VOLL system in place in the United Kingdom. Under the auspices of the Western Power Exchange (WEPEX), a voluntary design coalition including the three major investor-owned utilities (IOUs) and other stakeholders filed a market plan with the Federal Energy Regulatory Commission (FERC) in April 1996 and this filing envisioned active demand-side bidding. FERC provisionally approved the market model in November 1996[27] and approved a revised model in October 1997.[28]

AB 1890, the legislative enactment that enabled restructuring in California, was meanwhile signed into law in September 1996—at the same time the WEPEX plan was with FERC. AB 1890, however, contained a rate freeze that severely limited the role that demand-side bidding would play in the market for the duration of the rate freeze and placed restrictions on utilities' ability to contract forward for power. The motivation of the rate freeze was to provide guaranteed rate reductions to customers while at the same time recovering the stranded costs of the IOUs. The legislative process modified previous "floating cap" proposals and fixed the level of rates.[29]

In retrospect the provisions of AB 1890 in combination with the market structure created a recipe for exposure to unlimited wholesale power acquisition costs and wholesale power market prices. The rate freeze that limited the effectiveness of demand-side bidding gave rise to a nearly totally inelastic demand for wholesale power. The safeguard of demand bidding was eliminated without replacement by alternatives that would have served a similar purpose such as an installed capacity requirement.

When supply and demand were in reasonable balance the effect of the rate freeze, which limited the effectiveness of demand-side bidding, had little impact on the market. The summer of 2000 was the first real test of the market. Significant demand increases in load throughout the Western Systems Coordinating Council (WSCC)

region combined with hydro availability of the order of 15 percent below normal (in contrast, 1999 was of the order of 25 percent *above* normal) resulted in California exporting rather importing power and placed the state in a situation where capacity quite often was barely sufficient or insufficient to meet demand. With capacity approaching the level of demand plus operating reserve requirements, the demand inelasticity resulting from the rate freeze became a significant factor. Virtually all supply was required and competition among generators to produce and sell to the market was no longer a disciplining factor on market price. As all generators were required to meet California and export demand, the only practical limiting factor on price was the price caps instituted by the independent system operator. The demand response, which was originally envisioned as a governor on price, had been removed by the rate freeze. Programs to elicit a demand response were implemented but were not of sufficient scale to provide mitigation against extreme prices.

The California wholesale market has ultimately been subject to prices set at the whim of generators due to the inelastic nature of demand and the ability of generators to recognize market conditions under which nearly all supply will be required. Under these conditions, generators can have extreme pricing power without the need to withhold capacity to any visible degree.

California's problems are, of course, multitudinous and not just limited to the demand side. It has, for example, been severely constrained by the impossibly slow planning approval process for new generators,[30] by poor design of the trading arrangements,[31] and by other factors that increased price.[32] Southern California was subject to sudden and severe increases in environmental permit prices in 2000. Gas prices increased substantially in North America and more so proportionately in the West. There were important individual plant outages in the summer of 2000. Thermal units that were decades old and accustomed to running at low capacity factors were run flat out during that summer. When many inevitably required maintenance by the end of the year, their non-availability collectively led to further shortages in California in what is traditionally a shoulder-load time of year. During all of this, the California Market Monitoring Committee alleges that additionally, strategic bidding or market power might have led to further increases in price.[33]

It is not the intention of this section to delve into all of the factors here behind the disastrous events in California, except to conclude that the California market appears in many hours to have cleared at prices well above those that were expected. While limited programs were in place to elicit demand response, the rate freeze implemented by AB 1890 and the lack of general programs for enabling end users to respond to price means that the market cleared at prices that may have been well in excess of efficient prices that would prevail if a proper demand response system was in place. Further, the jump in prices in high-load hours may well be much greater than would exist in a market where the pricing systems and demand response programs had been implemented. The fact that the market often cleared at or near the price cap suggests that the market did not find the equilibrium price. That price, given the constraints of an unformed demand response and resulting inelastic demand, may be even higher than the price cap.

A discontinuous price curve with very high price spikes resulting from near-vertical demand and supply curves is an unusual phenomenon. While electricity may well by its very nature have such a situation, at least a part of the inelasticity is artificial in that end users do not see actual spot prices and have not prepared to react to those prices.

Endnotes:

1. Staff Report to the Federal Energy Regulatory Commission on the Bulk Power Markets in the United States, Nov. 1, 2000, Part II, at ii.

2. The market price signals will also be used to provide the correct incentives with respect to electricity *consumption* decisions.

3. During the transition to the new market model, the risk of unreasonable or unpredicted and sustained price spikes requires careful management. Transitional measures involving contracts that hedge both utility and consumer exposure to price spikes are required. Enormous utility or consumer exposure—what happened in California—must be avoided.

4. William W. Hogan, *Market Mechanisms and Decomposition for Coordination of Transmission Line Loading Relief Across Multiple Regions*, Harvard Electricity Policy Group, Milwaukee, WI, Nov. 20, 1998.

5. Of course, because of considerable uncertainties in estimating VOLL and LOLP, engineers tend to round off their calculations, and express them in politically digestible rule-of-thumb terms, for example, a target level of reliability such as one in 5 or 10 years.

6. "Oil on water" is an apt expression, used in Benjamin F. Hobbs, Javier Iñón, and Steven E. Stoft, *Installed Capacity Requirements and Price Caps: Oil on the Water, or Fuel on the Fire?*, ELEC. J., July 2001, at 23–34.

7. ISO New England is working on promising concepts of elastic demand curves for reserve (and hence price-sensitive ISO purchasing of reliability) that attempt to smooth the transition.

8. There are other obstacles in the transition from the vertically integrated model to the market model. Another source of inconsistency between the old model and the new lies in the conservatism of central-planning bureaucrats. Truth be told, when generation was regulated, utilities were spending other people's money and the more they spent the better. Regulators were not spending their own money either. Vertical integration endowed us with such gems as massively stranded nuclear costs and qualifying facility (QF) programs. The price duration curve (see Appendix II) that the market model inherits when it starts operations is at times massively distorted compared to economically efficient levels. For example, prices can start out very low because of these old, inefficiently capital-intensive plants, and the restructured market cannot cope in time with the radical changes that can follow: The initial low prices do not signal investment, but as new capacity becomes needed, prices suddenly spike to very high levels. The reason for the sudden price spike is that without the steady-state smooth price duration curve there is little in the way of price signals warning market participants of impeding shortages. By the time price signals arrive, it can be too late for the sluggish process of supply-side expansion to react. In the meantime, the politicians threaten to pull the plug on deregulation.

PII S1040-6190(01)00249-4

9. In the decentralized example of Appendix II, the fixed cost of the peaker was conveniently assumed to be zero. But what if the resource with the highest variable cost also has a fixed cost to recover? Something has to give.

10. The most expensive "resource" used to clear the market is no longer a generator. After all the generators are selected, the market is cleared by customers choosing not to consume. As the price goes up, more and more customers choose not to consume until the point where supply exactly equals demand. The price of electricity is set by the lowest $/MWh reservation price of the customers left consuming (so long as the pricing rule is correctly specified so as to let demand bids set the price). Because these reservation prices have no fixed cost component, electricity prices are consistent with reservation prices. Electricity prices are raised high enough so that the peaker capital cost problem is solved. (The analogy in the vertically integrated model is that a central planner would not have built capacity to serve these few customers that have a low value for energy consumed.) Potential new generators with fixed costs have prices against which to evaluate entry and make efficient decisions.

11. In contrast, when there isn't an active demand side, when supply is, or is nearly, exhausted, the inelastic nature of demand creates a situation in which the equilibrium price may be driven to well above the efficient price level.

12. Consumers and producers can hedge against price volatility and still face price incentives. Financial "contracts for differences" can still be arranged between any two individual market participants, for example, to reduce the price risk of each, while leaving the incentives to market participants in aggregate unchanged. Electric spot markets appear to be subject to extreme price volatility. While some of this volatility may be a function of the lack of mechanisms for eliciting a demand response and imperfect design and functioning of markets, it is also the case that spot electric prices are subject to significant variations as a result of relatively small imbalance between supply and demand. This means that while these markets may well provide efficient price signals, it is very risky for either consumers or producers to not hedge most of their exposure to these markets.

13. The New York Independent System Operator (NYISO) announced in a press release on Jan. 30, 2001, that it expects to make arrangements with customers representing only 200–300 MW of its 30,000-plus MW load to accept curtailment under its Emergency Demand Response Program (EDRP) plan in which such customers are *paid* the higher of $500/MWh or the prevailing real-time price, in the event they are curtailed.

14. As an alternative to representing load patterns over a month, load profiles can use broad peak and off-peak periods within a month if metering records usage this way.

15. There are many ways to establish the load profiles, some simple and some very complex. In some places all customers without hourly meters are assumed to have the same profile, which is equal to the difference between the distribution company's total load and the sum of the hourly metered loads. In other places, load profiles are prepared for many subsets of each customer class. Some systems set the load profiles ahead of time (sometimes with a weather variable that triggers use of a particular profile), while others determine the load profiles after the fact.

16. Demand response from load-profiled customers might be improved if willing customers' loads can be controlled directly as a substitute for customer response in high-cost hours.

17. Provided by Sally Hunt of NERA.

18. The old New York Power Pool at one time made plans to phase out installed reserve requirements as the amount of "dispatchable demand" increased sufficiently (personal recollection).

19. As an example of price response, Bushnell and Mansur reported that when retail electricity rates in the San Diego Gas & Electric service territory more than doubled over a period of three months in the summer of 2000, demand decreased by roughly 1.6% between mid-July and the end of August, and that demand was reduced by over 6% in the late afternoon and late evenings. That 1.6% might not sound like a lot, but it could have a significant impact on reliability and could grow over time as customers become

In some places all customers without hourly meters are assumed to have the same profile.

  PII S1040-6190(01)00249-4

more aware of price. James Bushnell and Erin Mansur, *The Impact of Retail Rate Deregulation on Electricity Consumption in San Diego*, Berkeley Power Conference, Berkeley, CA, March 2001. (Much of the increase in San Diego was subsequently rescinded through a retroactive rate freeze.)

20. Greek Electricity Law No. 2773, 1999.

21. Secretaría de Energía, *Propuesta de Cambio Estructural de la Industria Eléctrica en México*, México, 1999.

22. See also R. Turvey and D. Anderson, Electricity Economics: Essays and Case Studies (Baltimore, MD: Johns Hopkins University Press, 1977).

23. A represents the peaker plant, B represents the mid-load plant, and C represents the base-load plant.

24. The example in Figure 10 uses three types of plant, but in general the conceptual framework of this analysis can be applied to any number of different types of plant.

25. The slope of the demand response "peaker" will change, i.e., it will be a curve representing the different VOLLS of different consumers.

26. Parts of this description of the events in California are derived from Southern California Edison Company's description in its Opening Brief before the Public Utilities Commission of the State of California in Rulemaking 99-11-022 (Order Instituting Rulemaking into Implementation of Public Utilities Code Section 390), June 1, 2000.

27. Pacific Gas and Electric Co., San Diego Gas and Electric Co., and Southern California Edison Co., 77 FERC 61,204 (Nov. 26, 1996).

28. Pacific Gas and Electric Co., San Diego Gas and Electric Co., and Southern California Edison Co., 81 FERC 61,122 (Oct. 30, 1997)

29. The motivation for the contract restrictions was reported as being a desire to prevent "cream skimming," whereby large industrial buyers might allegedly be able to obtain contracts at more favorable prices than residential consumers.

30. States outside of California also have siting problems. On the first anniversary of operations, the NYISO Board of Directors, for example, identified inadequate generating capacity, the lack of price-sensitive load, and the absence of incentives for transmission expansion as the most significant threats to a robust, competitive market in New York State. Regarding new generating capacity, in an Oct. 19, 2000, press release, Board Chairman Richard J. Grossi stated the board's unanimous view that New York State's Article X generation siting process must be accelerated.

The NYISO Board is committed to competitive markets that will produce the intended benefits of deregulation, but the lengthy time required for the state approval process has undermined the market's ability to respond to the clear signals for more investment in generation. New Yorkers throughout the State must recognize that both conservation and expanded, environmentally benign, generating capacity are critical to maintain reliability and to ensure strong competition in the marketplace.

31. See, for example, William W. Hogan, *Designing Market Institutions for Electric Network Systems: Reforming the Reforms in New Zealand and the U.S.*, presentation made at The Utility Convention, Auckland, New Zealand, March 13, 2001.

32. Other market design flaws and the restriction on hedging compound the market flaw of inelastic demand in California. One of these other design flaws is manifest in the price differentials between California Power Exchange (CalPX) and ISO markets. Prior to the demise of the CalPX, most generator revenue from market sales was derived from the CalPX day-ahead market, as was the intention in the market design. The ISO market for imbalance energy was designed to account for the differences between day-ahead forecasts of load and supply and actual real-time consumption. The design of the markets, ever since the CPUC's original prescription, intended that through forces of market arbitrage, prices in real-time would converge to day-ahead PX prices. The actual outcome, however, a divergence in prices, pointed to an increase in costs, a loss of efficiency, and the existence of gaming opportunities.

33. From the point of view of the utilities, the California disaster was not just a problem of high prices, but also of *exposure* to those high prices. Two of the IOUs were still under the price freeze when prices rocketed upward. Furthermore, they were considerably more exposed to the high prices than they might have been otherwise: A requirement of the restructuring was that they divest 50 percent of their fossil fuel assets, so as to promote workable competition in wholesale generation markets. In practice, they ended up divesting much more. Their reputed reasons are complex and are said to involve their expected stranded cost recovery and otherwise allowed rate of return on nondivested assets—but regardless of the reason, the combined effect was to leave Southern California Edison Co. and Pacific Gas and Electric Co. highly exposed to wholesale electricity prices. If they had been permitted to hedge their exposure, coincident with divestiture, a much smaller portion of wholesale purchases and sales would be subject to market prices that FERC ultimately said were not just and reasonable. The practical damage would have been substantially mitigated.