regulation2point0.org

POWER

PWP-081

The Trouble With Electricity Markets (and some solutions)

Severin Borenstein

January, 2001

This paper is part of the working papers series of the Program on Workable Energy Regulation (POWER). POWER is a program of the University of California Energy Institute, a multicampus research unit of the University of California, located on the Berkeley campus.

University of California Energy Institute
2539 Channing Way
Berkeley, California 94720-5180
www.ucei.org

The Trouble With Electricity Markets (and some solutions)

Severin Borenstein*

January 2001

Abstract: Since June 2000, California's electricity market has produced extremely high prices and threats of supply shortages. But California's is only the most recent and visible example of the trouble with deregulated wholesale electricity markets. In this paper, I argue that the difficulties that have appeared in California and elsewhere are intrinsic to the design of current electricity markets, in which demand exhibits virtually no price responsiveness and supply faces strict production constraints. Such a structure will necessarily lead to periods of shortage and of surplus, which will be accompanied by great volatility in prices and profits. This result, however, is not inevitable. By encouraging price-responsive demand and long-term wholesale contracts for electricity, policy makers can create electricity markets that will function much more smoothly.

* Director, University of California Energy Institute (http://www.ucei.org) and E.T. Grether Professor of Business Administration and Public Policy, Economic Analysis and Policy Group, Haas School of Business, University of California, Berkeley, CA 94720-1900. http://haas.berkeley.edu/~borenste. I have benefitted a great deal from discussions with Carl Blumstein, James Bushnell, Paul Joskow, Steve Puller, Steve Stoft, Frank Wolak, Catherine Wolfram, and Hal Varian, but the opinions in this paper do not necessarily reflect their views. Erin Mansur provided excellent research assistance.

I. Introduction

In much of 2000, California has seen unprecedented wholesale electricity prices. These high prices have produced enormous profits for generating companies and financial crises for the utilities that must buy power in the wholesale markets and sell at much lower regulated prices in the retail markets. Accusations of price gouging and collusion among the sellers have been widespread. Some observers have blamed the problems on the format of the wholesale auctions in California, while others have focused on the way that transmission capacity is priced and locational prices are set. A number of economists, myself included, have done studies that have concluded that sellers are exercising significant market power, at times raising prices to well above competitive levels.

In this paper, I show that while some or all of these issues may play a role in the difficulties California and other electricity markets have encountered, the policy discussion thus far has not confronted directly the fundamental problem with electricity markets: In nearly all electricity markets, demand is almost completely insensitive to price fluctuations and supply faces binding constraints at peak times. Combined with the fact that unregulated prices for homogeneous goods almost always clear at a uniform (or near uniform) price for all sellers – regardless of their costs of production – these attributes necessarily imply that short-term prices for electricity are going to be extremely volatile. Problems with market power and imperfect locational pricing only exacerbate the fundamental trouble with electricity markets.

Two policies would greatly mitigate the fundamental trouble: long-term wholesale contracts between buyers and sellers and real-time retail pricing of electricity, which will indicate to the final customer when electricity is more or less costly to consume.[1] Historically, long-term contracts have been the politically more attractive approach, with cost-of-service regulation being the most detailed and extreme form of long-term contracting.[2] Long-term contracts allow buyers to hedge against price booms and sellers to hedge against

[1] This prescription is certainly not novel, see Joskow (2000), but contracts have only recently been accepted by regulators in California and real-time retail pricing still is not a priority in most restructured markets.

[2] Cost-of-service regulation is effectively a cost-plus contract between the regulated, vertically-integrated utility and the customers, where payment is adjusted for virtually all changes in the utility's cost of production. Defense contracting has often been conducted in this way.

price busts. The long-term contracts now being considered would engender more performance incentives than traditional regulation. The simplest form is one that simply sets a price and quantity to be delivered at every point in time, and leaves it to the producer to try to increase its profits by meeting that supply commitment in its most cost-efficient manner.

While long-term contracts surely must be part of the solution, they should not be used as the entire solution. A much more cost-efficient and environmentally responsible response to the problem combines long-term contracting with real-time retail pricing. At first glance, it may seem that long-term contracting and real-time retail pricing are not compatible with one another. I show in this paper, however, that prices can reflect real-time variation in the price of electricity while monthly electricity bills can remain quite stable through the use of long-term contracts. Furthermore, implementing real-time retail pricing could substantially reduce the prices buyers would need to offer to procure long-term contracts. Together, these two policy responses would help to produce an electricity market that operates in a smooth, cost-effective, and environmentally responsible manner.

II. Why Are Electricity Prices So Volatile?

It is by now well-recognized that the physical properties of electricity production make the matching of supply and demand especially difficult, while the physical properties of electricity transmission and distribution make it critical that supply exactly match demand at every moment. Because storage of electricity is extremely costly and capacity constraints on production from a plant cannot be breached for significant periods without risk of costly damage, there are fairly hard constraints on the amount of electricity that can be delivered at any point in time. Yet, because of the properties of electricity transmission, an imbalance of supply and demand at any one location on an electricity grid can threaten the stability of the entire grid. The supply/demand matching between any customer and supplier is just part of the overall grid balancing and any mismatch could disrupt delivery of the product for *all* suppliers and consumers on the grid.

Given these unusual aspects of the supply side of the market, it is all the more remarkable how little flexibility has been accommodated on the demand side of the market. While the technology to meter consumption on an hourly, or even 10-minute, basis is widely available, and has even been installed at many industrial and commercial locations, no electricity market in operation today makes substantial use of real-time pricing, *i.e.,* charges a customer time-varying prices that reflect the time-varying cost of procuring electricity at the wholesale level.

A third aspect of the electricity industry completes the circumstances that lead to volatile prices: electricity generation is very capital intensive. Because a significant part of generation costs are fixed, the marginal cost of production will be below the average cost for a plant operating at below its capacity. So long as the market price is above a plant's marginal operating cost, a competitive firm is better off generating than not. As a result, excess capacity in a competitive market will cause prices to fall to a level below the average cost of producing electricity, and generators will lose money.[3] This capital intensity, implying a high cost of idle capacity, is also the reason that it is very costly for firms to maintain the ability to increase electricity production on very short notice.

The effect of these characteristics of the electricity market can be shown graphically. Figure 1 illustrates a situation in which very-inelastic supply and very-inelastic demand intersect at a price that permits producers to cover their capital costs. This is a price that allows the plant to cover its marginal operating costs and earn enough beyond that to justify its capital investment, both depreciation and return on investment. But it is easy to see that if capacity cannot adjust quickly and demand is difficult to forecast precisely, figure 1 is an unlikely outcome. Even small changes will lead to a price boom or bust.

Figure 2 illustrates a slight mismatch in which quantity demanded exceeds supply at normal price levels. Unlike, for instance, in the airline industry, where capacity on a route can adjust quickly and demand is responsive to price changes, there is almost no price-responsive mechanism on the supply or demand side that allows the electricity market to adjust to such a mismatch. In this situation, price will skyrocket, eventually eliciting a bit more output as generators run their plants harder – risking heavier maintenance costs – due to the tremendous profit opportunity. Extremely high price can also elicit

[3] This potential has been quite apparent in other capital-intensive industries, for instance, the memory chip market. In the early 1990s, high prices induced massive investment in memory chip fabrication plants. The resulting excess capacity caused prices of memory chips to collapse and producers to lose billions of dollars.

demand response, but in the current markets, this is quite limited. The most prevalent sources of "demand responsiveness" are attributable to actions by the independent system operator (ISO), which can reduce reserve margins and can exercise interruptible contracts, an extreme measure that causes significant disruption to the affected loads and is thus one that ISOs are reluctant to take.

The situation in figure 2 is exacerbated if markets are not completely competitive. As I and others have demonstrated elsewhere,[4] tight supply conditions in electricity markets put sellers in a very strong position to exercise market power, raising price above the level at which competitive supply and demand would otherwise meet. Because market power is easier to exercise in electricity markets when the competitive price would have been high anyway, it raises prices more during demand peaks that during off-peak periods. Thus, the presence of market power exacerbates the volatility of prices and further reduces the chance that prices will remain in a reasonable range.

Many observers of deregulation have said that the root of the problem in California is that the surplus of capacity that was supposed to exist in California disappeared due to strong economic growth in the state and throughout the western U.S. If the surplus had remained, however, the result would have been figure 3. Figure 3 illustrates a mismatch of supply and demand in the opposite direction, with supply exceeding demand at normal prices. With excess supply, price is likely to collapse to the low marginal running costs of the marginal unit. These prices would almost certainly fail to cover the average costs of operating the plants, a situation similar to the memory chip market in the early 1990s. Experience from many other industries – oil, dairy, or wheat, for instance – supports the political reality that such outcomes are frequently followed by calls for price supports and subsidies to producers, essentially reregulation due to excessively low prices.

The careful reader will note that figures 1-3 each illustrate one supply/demand interaction, while in real electricity markets demand varies continuously throughout a day or month, and prices are set hourly or more frequently. The concept illustrated in these figures, however, also applies to the aggregate of these interactions. Consider, for in-

---

[4] See Wolfram (1999), Borenstein & Bushnell (1999), Borenstein (2000), Borenstein, Bushnell & Wolak (2000), and Joskow & Kahn (2000).

stance, the supply curve shown in figure 4.[5] Begin by assuming that demand in a month is distributed uniformly between D_L^a and D_H^a. Now, consider a relatively small rightward shift of the demand distribution to between D_L^b and D_H^b. This small shift replaces hours that were at very low prices, on the left of the distribution, with hours that are at extremely high prices at the right side of the distribution. Even this small shift can cause the quantity-weighted average price to increase drastically.

The critical point here is that electricity markets are especially vulnerable to these supply/demand mismatches due to the extreme inelasticity of supply and demand. In markets where output is storable or capacity constraints are more flexible, supply can adjust to such mismatches without extreme price movements. In markets where buyers can see time-varying prices and respond to them, demand can adjust to such mismatches and thus damp the price swings. As currently configured, electricity markets have almost no flexibility on either the demand or supply side. California is currently in an electricity price boom, but if the state had suffered a significant economic downturn in the late 1990s – which was a real possibility in 1994 when the state began serious consideration of wholesale price deregulation – the opposite outcome could have occurred. The price volatility problem is exacerbated by the exercise of market power, raising prices above the outcome that would occur with the competitive supply and demand that is illustrated in figures 1-3, but doing so to a much greater extent at times when even competitive prices would have been very high.

A comparison of electricity markets to the airline industry is potentially useful, because that is also an industry with (a) time-varying demand that is difficult to forecast, (b) strict short-run capacity constraints, and (c) an inability to store output (*i.e.,* once a plane takes off, any empty seats are pure waste). The key difference is that retail prices in the airline industry vary with demand – it is very hard to find a discount seat on a Friday afternoon at 5pm – and customers are aware of the prices, and respond to them, at the time they choose to purchase. Furthermore, mismatches between supply and demand on a route, or even in a region, can be remedied in a matter of days by flying in more capacity. Capacity expansion

[5] One could think of this as a competitive supply curve if the market is competitive. If firms have the ability to exercise market power, one could think of this roughly as a "quantity/price outcome" curve that also incorporates prices that deviate from marginal cost.

in the electricity industry takes years. Electrical transmission lines can substitute to some extent for production capacity at a location, making it easier to bring in power produced elsewhere, but those transmission lines also have capacity limits and new lines also take years to build.[6]

Understanding Price Formation in a Commodity Market

The discussion so far has assumed that all sellers in a short-term market for electricity receive the same price for delivery of power at the same time. There has been a great deal of discussion about the fact that sellers who show a willingness to sell at low prices are paid a much higher market-clearing price. This is, however, the way that all commodity markets work. If an orange grower has a particularly fertile orchard that produces high yields and low costs per bushel, that grower still sells oranges at the market-clearing price. Likewise, gold mining companies and oil producers sell their output at the market price regardless of whether they are producing from low-cost or high-cost sources.[7]

This uniform-price outcome is not a function of the auction format or some design flaw in the electricity market. Furthermore, it is true in all commodity markets, whether or not there are firms that are able to exercise market power. This reality, however, has important implications in electricity markets due to the extreme volatility of prices, as I've discussed above. When a supply/demand mismatch occurs, it changes the price for all power being sold in the market at that time.

The uniform-price outcome is in sharp contrast to the outcome under regulation, in which price is based on the *average* cost of production rather than the *marginal* cost (or marginal opportunity cost) of production. In other words, under regulation, production that takes place at lower costs is compensated at a lower level. The price consumers pay is the average of all of these production costs. Assuming that production is equally efficient

[6] Transmission losses, which can dissipate as much as 10%-20% of the marginal quantity transmitted, increase the marginal cost of this power even before hitting the capacity constraint. But if a cost increase of that magnitude were the only barrier to bringing in ample supply, we still would not see 10- or 100-fold price spikes.

[7] The same idea applies even to housing markets, where products are differentiated. The price most sellers are willing to accept is based on current market conditions, and is unrelated to the owner's "cost" as indicated by the price at which she bought the house or the length of time for which she owned it.

under regulation or in a competitive market,[8] average cost pricing will yield lower prices when supply is tight, because the marginal cost of production is above the average cost. The difference will be even greater if the unregulated market is not completely competitive and unregulated prices are above marginal cost. That is the situation California currently faces. But in a situation of surplus capacity, marginal cost will be below average cost. So, unless the markets suffer from significant market power, the price that results from a market process will be below the price that regulation would produce. That is the situation that many people believed California would face in the early years of restructuring.

III. The Role of Long-Term Contracting

In unregulated markets that exhibit a great deal of spot-price volatility, it is common for buyers and sellers to smooth their transaction prices by signing long-term contracts. This can and should be a significant part of any electricity market undergoing deregulation. Nearly all markets outside of California have taken this approach. In many cases, the sale of utility generation facilities to other firms has been accompanied by "vesting contracts" that promise a certain amount of power sales back to the utility at a predetermined price. Also, the regulated utilities, now operating primarily as utility distribution companies (UDCs) and electricity retailers, have in many cases retained some of their generation facilities. The price customers end up paying for the power from those facilities is then based on the costs of operating those facilities, not the market price. While California had virtually no vesting contracts, the California utilities have retained generation facilities that produce nearly half of the power they deliver to customers. If they had sold all of their generation facilities, the situation in California would now be much worse than it is.

Some participants in the debate have suggested that utilities in California and elsewhere could save money by purchasing their power through long-term forward contracts.[9] While long-term contracts reduce variability in the cost of buying power, long-term contract prices are unlikely to be systematically below spot prices *on average*. In the case

[8] This is an assumption with which many people would take issue.

[9] The discussion here refers to long-term contracts that are negotiated in a market context, whether bilaterally or in a more organized forward market. Cost-based long-term contracts that are imposed through a regulatory process offer much of the same hedging benefit, but the relationship between long-term contract and spot prices is not as clear cut.

of summer 2000 in California, power contracted in advance was cheaper than spot power, but the reason sellers were willing to contract at those lower prices when the contracts were signed in late 1999 or early 2000 was that their best guess of summer 2000 prices was below the spot prices that actually resulted. The offer prices for power over the next few years in California are quite high right now because sellers currently anticipate that spot prices over that time period will be quite high.[10] In the Pennsylvania-New Jersey-Maryland pool, by contrast, buyers who contracted for summer 2000 power in advance ended up paying higher prices than those who bought on the spot market, because mild weather led to prices that were below expectations. *On average, a purchaser buying power in forward markets (or through long-term bilateral contracts) will not receive lower power costs than a purchaser buying in the spot market.*[11]

There is, however, a potential price-lowering effect in both forward and spot markets if buyers in aggregate purchase more power in long-term contracts. Locking in some purchases in advance reduces the ability of multiple firms to behave less competitively among themselves.[12] The idea, basically, is that if firms are maintaining high prices by mutually foregoing aggressive price cutting, then the existence of many forums for trading, especially over time, makes it more difficult to maintain such mutual forbearance.[13] Thus, the possibility of selling in advance makes it more difficult for firms to restrain competition, in much the same way that the possibility of secret price cutting has this effect. Once a firm has sold some output in advance, it has less incentive to restrict its output in the spot market in an attempt to push up prices in that market.[14] Thus, in anticipation of more

---

[10] Statements by various parties that trumpet the savings buyers obtain by purchasing power through forward markets do not mention the losses of exactly equal size that producers experience by *selling* in advance. Generators would not be likely to continue to sell in the forward market if they knew they could systematically earn more in the spot market.

[11] Borenstein, Bushnell, Knittel, and Wolfram (BBKW, 2000) looks at the relationship between the price in the California Power Exchanges day-ahead price and the California ISO's balancing market price. As discussed by BBKW, there could be a systematically lower price in the forward market if sellers are systematically more risk averse than buyers, but this is unlikely to occur.

[12] This insight is generally attributed to Allaz and Vila (1993).

[13] The mutual forbearance can take the form of implicit or explicit collusion or it can be firms simply competing less aggresively than they could by, for instance, limiting the quantity of power they make available to the market (called "Cournot competition").

[14] Borenstein (2000) discusses this.

aggressive competition in the spot market – because some firms have presold significant output in a forward market – firms are likely to price more aggressively in the forward market. The actual magnitude of this potential impact of forward contracting on both spot and forward market prices is not known. It can do no more than eliminate the portion of price premia that are due to market power, and it might have a substantially smaller effect. Forward contracting cannot reduce the price increases that would accompany supply shortages in even a completely competitive market.

From a buyer's perspective, the danger of long-term forward contracting, of course, is that the buying firm might lock in a higher price than it would have had to pay if it had purchased in nearer-term markets. This is an especially large fear for utilities acting as energy service providers (ESPs) in a restructured market. They are concerned that in such a situation the state regulatory agency might decide that the contract purchase price was "imprudent" and not allow the utility to pass through the costs to customers. Credible commitment by regulators is difficult, but it is clear that the correct standard for judging the prudence of these contracts is based on the information available to all parties at the time the contract is signed, not looking backwards after the actual spot prices have become available. A hedging instrument is not going to save money on average, and will frequently lose money, compared to the relevant spot prices.

Long-term contracting is an important part of the solution to the fundamental problem of electricity markets, but by itself it is not the most efficient solution. Long-term contracting does not "solve" the mismatches between supply and demand. It just prevents large fluctuations in electric bills when those mismatches occur. It can, however, help pay for excess or stand-by capacity, by assuring that the generating companies receive payment of their capital costs even if demand turns out to be low, spot prices collapse, and some of the capacity does not get used.

In fact, this is what the old regulatory system did. Utilities were assured of revenues to cover their costs and in return built sufficient capacity to make sure that all contingencies could be covered. Supply always exceeded demand by a significant amount, and the cost of all that idle or reserve capacity was rolled into the price that customers paid for the power that they did use.

This system could still work in a quasi-deregulated form in restructured electricity markets. Utilities (or other ESPs) could, with the oversight and consent of the regulator (or without it, for unregulated ESPs), sign long-term contracts for power and capacity (or simply take-or-pay contracts) that assured generators they could recover their costs even if the capacity were not actually used. This would mimic the regulatory system and would produce many of the same inefficiencies.

This would be an unfortunate outcome. In particular, it would still fail to bring the demand side into the market. Demand-side price responsiveness is valuable in this market because it does not make sense to produce more power if that production imposes costs that are greater than the customer's value of consuming the additional power. Real-time retail prices that reflect the cost imposed by additional consumption are the ideal mechanism for making that tradeoff.

Thus far, California and all other states have attempted to make electricity markets work almost entirely on the supply side of the market. At times, this has worked relatively well in some markets, but recent events in California have demonstrated the variety of constraints that exist on the supply side. In California, these include production capacity constraints, new plant siting constraints, pollution emission constraints, and constraints on the quantity of natural gas that can be shipped into the state in any given time period. Deregulating only the supply side of the market seems to be the equivalent of making an electricity market operate with one arm tied behind its back.

The complete solution to the fundamental trouble with electricity markets includes a combination of long-term contracting for supply and real-time retail pricing for customers. Together, these mechanisms can provide the right economic incentives to reduce demand at peak times when the system is strained while still assuring customers of relatively stable monthly bills.

IV: Real-time Retail Price Signals and Stable Monthly Bills

A distinct feature of electricity markets in the U.S. is that although the marginal cost of delivering the product can vary tremendously over time and hard capacity constraints are present, retail prices are seldom adjusted to reflect these cost (or opportunity cost) variations.[15] The effect of customers facing a single constant price for electricity is that they have no more incentive to conserve during peak consumption times, such as on a hot summer afternoon, than during low consumption times, such as during the night. They also have no incentive to shift consumption away from times when the production capacity of the grid is strained and production costs are highest. As a result, more capacity needs to be built to accommodate all of the demand at the highest peak times than would otherwise be the case. Real-time pricing would reduce the need to site and build new peaking plants. If consumption is much less price-elastic at low prices than at high prices, then real-time pricing is also likely to lower the overall consumption of electricity. Thus, in addition to the potential cost savings for customers, the environmental impact of real-time pricing makes it quite attractive.

While many people have advocated greater price-responsiveness in demand through real-time retail electricity pricing, at the same time, there have been calls for greater protection of customers from price spikes. These two positions are inherently incompatible since price-responsive demand will take place only when customers are exposed to variable prices. But the underlying goals of these two policy proposals are not incompatible. It is possible to expose customers to nearly the full range of price fluctuations, so that price-responsive demand will be meaningful, and still assure them of relative stability in their monthly bills. It seems clear that customers are concerned about stability in their monthly bills, not about stability in their hour-to-hour prices.

The key to meeting both of these goals is to recognize that the overall or *average level* of prices can be stabilized without damping the variation in prices. In order for an energy service provider to offer both real-time retail price variation and monthly bill stability, without risking substantial losses, it needs to hedge a significant portion of its energy cost through long-term contracts. With a large share of power purchased under long-term contracts, the ESP can adjust all of the hourly real-time retail prices up or down as necessary each month in order for the payments under real-time pricing to equal the

[15] Outside the U.S., peak/off-peak pricing is much more common, though it is virtually always implemented as a two- or three-price system with, for instance, one price for daytime usage and a lower price for nighttime usage. Real-time retail pricing, in contrast, allows prices to change with each given time interval, such as 10 minutes or one hour, and prices need not be the same at a given time from one day to the next.

costs of procurement.

To be concrete, assume that the ESP begins by engaging in no hedging. It charges customers a fixed per-kilowatt-hour transmission and distribution (T&D) charge plus the spot price of energy in each hour.[16] This satisfies the real-time pricing goal, but does not assure stable monthly bills. In fact, the monthly bills would be as variable as the month-to-month variation in the weighted-average spot energy prices.

To attain the goal of monthly bill stability, the ESP would sign a long-term contract to buy some amount of power at a fixed price.[17] To fix ideas and keep the presentation simple, I'll assume that the ESP signs a long-term contract at the same price for each hour. Such a contract is likely to be at about the average spot price of the electricity that the parties anticipate over the life of the contract, but in any given month the contract price could be greater or less than the average spot price.

This contract can be considered a financial investment that is completely independent of the ESP's retailing function. The critical point is that the ESP's return on this financial investment varies directly with the average spot price of energy, and that return can be applied to change the average level of customer bills. When viewed this way, it becomes clear that the long-term contract can affect the average price level without damping the price variation. The gains (when the average spot price is higher than the contract price) or losses (when the average spot price is lower than the contract price) from the long-term contract would be distributed to customers by a constant (over the month) surcharge or discount on each kilowatt-hour sold during that month.

After the fact, the ESP would then charge the customer the spot price plus or minus the return per kilowatt-hour from the long-term contract. Since the return to owning the contract is greater when the spot price is higher, it would be used to offset the high average spot price, thus lowering the volatility of monthly electricity bills. If the ESP hedged nearly

---

[16] The spot price here could refer to the PX day-ahead price or the ISO imbalance energy price. Obviously, this entire discussion assumes that the customer has installed a real-time electricity meter. The cost of such meters is not a significant expense for most commercial and industrial customers, but could be for many residential customers.

[17] In fact, the contract just has to have less variance than the spot price. It could, for instance, have a fuel adjustment clause.

all of the load it served, then this offset would be sufficient to nearly eliminate variability in monthly bills. If the ESP hedged, say, 80% of the load, then about 80% of the variability in monthly bills would be eliminated. The actual price that the customer was charged in each hour, however, would still have the same variance as the spot price.

During the month, customers would not know the actual price they will be charged for each hour, but they would be able to learn the spot price at any point in time, so they would know in which hours electricity will end up being extremely expensive and in which hours it will be much cheaper. The offset, since it is averaged over all hours, would be just a few cents per kilowatt-hour, so the hedging of the average price would not interfere with the "passthrough" of real-time price signals.

Figure 5 illustrates the effect that this approach would have had in June of 2000. In this illustration, the utility is assumed to have signed a contract for 80% of its load at 6¢ per kilowatt-hour. In addition to energy charges, the utility is assumed to assess a 4¢ per kilowatt-hour charge for transmission and distribution. The T&D charge is added to all prices for ease of comparison. The three horizontal lines show the load-weighted average price a customer would pay (assuming it had the same load profile as the system as a whole) if the utility was fully hedged (lowest line), if it was completely unhedged (highest line), and if it was 80% hedged (middle line).

Of the two volatile lines, the higher shows the real-time price a customer would pay with no hedging and the lower shows the price the customer would pay under the approach proposed here that combines real-time pricing with 80%, in this case, purchased through long-term contracts. The load-weighted average of the higher line is 18.08 cents, the same as the highest horizontal line. The load-weighted average of the lower line is 11.62 cents, the same as the middle horizontal line.[18]

This illustration demonstrates that a customer under the plan proposed here would face the same volatility in prices as it would under 100% real-time pricing. The only difference is that the price curve would be shifted down, in this example, by the "profits"

---

[18] Though it does not occur in this illustration, it is possible that this formula could result in negative prices in certain hours. This outcome could easily be avoided, however, with a small modification. A minimum price, say 1¢ per kilowatt-hour, could be set and any resulting excess revenue could then be redistributed evenly among all other hours.

from the long-term contract, which is this example are 6.46¢ per kilowatt-hour. During the hours of extremely high spot prices, customers would face nearly as extreme prices, and would have a strong incentive to reduce consumption. Yet, the average monthly prices (and monthly bills) the customer would face would be much less volatile than without hedging.[19]

The most important impact of this approach would be that it would lower quantities demanded at peak times, and by doing so, it would lower the market prices at those times. Harkening back to figures 1-3, the demand curves would become much flatter, since customers would be able to see and respond to high prices, so the response would prevent the extreme price spikes that we now see. In doing this, real-time retail pricing would lower the overall average cost of power.[20]

This has very important implications for the negotiation of long-term contracts. If sellers, at the time of negotiation, believe that effective real-time retail pricing is likely to be implemented, then they will reduce their forecasts of the average spot prices they would be able to earn if they did not sell through a long-term contract. As a result, the sellers will be willing to accept a lower long-term contract price than they otherwise would. Thus, it would be quite valuable for the regulator (or ESP) to make a credible commitment to implementing real-time retail pricing before the ESP negotiates these long-term contracts.

Though real-time retail pricing has not been widely used in the U.S., the technology for it is well-established and available. Large commercial and industrial customers in California have real-time meters already, and communication of the day-ahead or balancing market price to those customers easily can take place through the internet. In the near

---

[19] This illustration slightly overstates the monthly bill stability that could be achieved through 80% hedging because it assumes that the hedged quantity is 80% of the actual demand in each hour. The contract (or contracts) would quite likely hedge a larger quantity during periods when demand is anticipated to be high, but the variation would probably not match exactly the actual variation in consumption that occurs. Since price will be highest in periods when the quantity exceeds anticipated levels, the protection from the hedging contract would be slightly less than if it matched the actual consumption pattern exactly.

[20] It is also worth noting that setting retail prices below the sum of the wholesale price and the transmission and distribution charge can move prices closer to the actual marginal cost, even if there is no market power present. Transmission and distribution is charged on a marginal basis, but these costs are largely fixed. Therefore, reducing price by up to the T&D fee that would otherwise be in the retail price has the effect of moving price closer to marginal cost.

future, it may not be practical or necessary to include residential customers in a real-time retail pricing program, but as the cost of real-time meters declines, there is no reason that it shouldn't occur. It is critical to understand that the *variation* in prices can be separated from the *average level* of prices. For any given level of flat retail price that is contemplated, the same average price level can be attained each month with real-time retail pricing. Doing it with real-time pricing will reduce the cost of procuring the power and reduce the need to build more power plants.

How Much Power Should be Bought Under Long-Term Contracts?

In the foregoing example, I assumed that the ESP signed long-term contracts for 80% of its power needs. That number represents a very significant share of total consumption, so it provides a great deal of stability in the cost of procuring power. Higher figures would provide even greater stability. However, as the ESP contracts for a larger share of its expected needs, it becomes more likely that if spot power prices drop below the levels anticipated, enough final customers might abandon the ESP (in a world with retail choice) that the ESP finds itself more than 100% hedged.[21]

If the ESP is more than 100% hedged at prices that end up exceeding the spot price of electricity, then it effectively has a stranded investment in long-term electricity contracts. If it is a private, unregulated ESP, then its shareholders will simply have to absorb that loss. If the ESP is the regulated utility distribution company, however, then it will likely ask the regulator to be allowed to recover that loss through either an exit fee on customers wishing to switch to another ESP or an uplift charge on those that remain. Either of these outcomes would be politically quite contentious.[22]

For these reasons, it seems wise for an ESP, especially one that is the incumbent UDC,

[21] This is especially likely if, as now seems contemplated in California, the UDC signs long-term contracts at "levelized" rates in order to smooth (or hide) the generators' collection of profits that they would otherwise earn mostly in the next year or two. Because the price of natural gas is expected to decline and new plants will ease the capacity shortage in California, a levelized long-term contract leaves prices higher than the expected spot price in more distant future years.

[22] From an economic point of view, an exit fee does seem to be the best way to deal with this possibility. The correct exit fee would be the customer's pro rata share of the expected loss on the remainder of outstanding long-term contracts. This would result in an economic incentive to leave the incumbent UDC only if the customer's new ESP can offer, *going forward*, to provide power at lower cost or with higher quality of associated services.

to hedge less than 100% of its anticipated power purchases through long-term contracts. Given the experience with long-distance telephone, even large commercial customers might not be very quick to abandon the incumbent UDC, so a large proportion of hedging still seems relatively safe.

Finally, as pointed out earlier, ownership of generation facilities is a form of partial hedging. To the extent that price movements are related to the scarcity rents associated with capacity or variations in the degree of market power that is exercised, owning capacity is an effective hedge against these changes. Capacity is not an effective hedge if price changes are due to changes in the cost of fuels and these are the same fuels that the retained generation uses.[23]

The Difference Between Real-Time Retail Pricing and Paying for "Negawatts"

Many alternative programs have been proposed that mimic, to some extent, the effect of real-time pricing. These programs generally are based on the idea of paying customers to reduce consumption at certain times. Interruptible loads are the bluntest instrument of this type. Other proposals suggest that the ISO should have a standard offer or run an auction of some sort for load reduction at the time the system is under strain.

While these programs can, at their best, offer many of the benefits of real-time pricing, in practice they offer less benefit and about the same cost as real-time pricing. On the cost side, with the exception of the very bluntest load interruption, implementing any sort of load-reduction market requires the same real-time metering equipment and about as much price, load, or reserve-margin information transmission as real-time pricing.

The more difficult problem with paying for load reduction is the baseline from which the payment is made. Unless the program is mandatory and the baseline is set based on information that is completely out of the control of the customer (such as load information from a number of years earlier), the program will be subject to extensive manipulation and self-selection problems. The manipulation occurs if the baseline is set based on any consumption information that can be affected after the program is announced or anticipated.

[23] Signing a long-term contract for the fuel, however, can then be an effective hedge. Hydroelectric power provides a varying and unpredictable degree of hedging year-to-year.

For instance, one current suggestion in California would pay a customer on superpeak summer days to reduce load from its average level over the previous x days. This would greatly diminish any incentive to reduce load in other days, since such actions would lower the baseline the customer started from on the superpeak days. The self-selection problem exists even if the baseline is set from truly exogenous information. The entities that would opt to sign up for these programs will disproportionately be the ones who know already that they will be reducing their loads, such as companies that are reducing their operations or that have already changed their production process to use less power.[24] Likewise, those entities whose baseline has been set inordinately high, due to some unusual activity during the period used for determining the baseline, would also be more likely to join the program.

Paying for load reduction at peak demand times seems, at first, more attractive than real-time pricing, because it "rewards" those who conserve at peak times rather than "punishes" those who consume when the system is strained. There are, however, no free lunches, and this distinction is a false one. Those payments to entities that conserve at peak times must come from somewhere. If the revenue source is an explicit or implicit surcharge on all power consumed – probably the most likely scenario – then prices are being increased for all other users, and the punishment is just as great for those who consume in the middle of the night as for those who consume during a hot afternoon. If the revenue source is general state funds or some other tax source, then it is coming from all users and the punishment is allocated in a way that has nothing to do with energy usage at all.

Finally, the payments in many of these programs fail to recognize the savings that a customer already gains by reducing consumption, the price they would have paid for the power. Thus, if additional power would cost $300/MWh to procure and the customer is paying $80/MWh, the maximum payment the ISO should be willing to make to the customer for load reduction is $220/MWh, not $300/MWh. From the provider's side, this

[24] A very similar self-selection problem occurs if real-time pricing is implemented on a voluntary basis. Those entities that know they consume disproportionately at the peak times will not opt for the program and will thus continue to have no incentive to conserve when the system is strained. Particularly for industrial and commercial consumers, real-time pricing is best implemented as the default pricing system. If a company then wanted to sign a contract with a power marketer to obtain flat prices, they could do so. Marketers, however, would sell such contracts at a very high price, because they would recognize that the buyers are customers who disproportionately consume at high-cost times.

is because reducing load saves procurement costs of $300, but lowers revenue by $80 for a net savings of $220. From the customer's side, it is effectively being "paid" $220 plus the savings of $80 for a total of $300 for every MWh it reduces.

Fairness and Distributional Concerns with Real-Time Retail Pricing

Real-time pricing would almost certainly lower the total payments for power by discouraging peak-time consumption that otherwise causes prices to skyrocket. Not everyone would benefit equally from a switch to real-time pricing, however, and some customers would be made worse off. Those who now consume disproportionately at times when the system load is highest could be made worse off. Under the current flat pricing of electricity, these entities are being subsidized by those that consume a smaller share of the system load at peak times than at off-peak times. In the case of California, this cross-subsidy roughly runs geographically from areas that do less air conditioning to areas that do more. There is concern that central valley communities could be harmed relative to coastal communities.

It is worth pointing out first that with even a moderate amount of price responsiveness, it is possible that nearly all customers could be made better off, as the real-time retail pricing reduces demand at peak times and prevents most or all of the extreme price shocks. Even though the cross-subsidy of peak-time consumption would be ending, the wholesale price at peak times would be reduced as demand at those times declines, so the increase in the retail price at peak times relative to flat retail pricing would not be nearly as great as one would infer from looking at recent price patterns. To the extent that policy makers wish to continue to cross-subsidize towards areas that consume more power at peak times, this could be done through an explicit and transparent subsidy of power use in those areas, preferably one that does not continue to subsidize consumption at peak times most heavily.[25] In the end, however, the only way to absolutely assure that *no one* will be made worse off by ending this cross-subsidy is to continue with flat pricing, which gives no incentive to reduce peak-time consumption.

[25] For public health reasons, one might want to exempt, or explicitly subsidize, low-income consumers who are reliant on electricity use at peaks times.

V. The Role of Price Caps

In California and other electricity markets, price spikes have led to price caps, and to a debate about the appropriateness of imposing price caps. Some participants are “for” price caps and others are “against” them. These philosophical debates about price caps are an unfortunate distraction from the real issues.

Price caps are and will continue to be a critical element of wholesale electricity markets. The markets as now configured do not give customers an opportunity to respond to high prices by reducing their consumption. Furthermore, the extreme inelasticity of both supply and demand mean that there is the opportunity for exercise of extreme market power, potentially driving prices to a thousand or a million times higher that their normal level. Such outcomes would destroy the market. Therefore, the debate should be about the level of price caps and mechanisms for their adjustment.

Those arguing against price caps have said that they will reduce investment in production facilities and reduce production from facilities that already exist. Both statements are potentially true. *If price caps are set too low, they will have detrimental effects.* Virtually no one disputes this. The question is at what level these effects will occur. The answer in economic terms is straightforward.

In the short run, a price cap will deter production from an existing facility if the cap is below the short-run marginal cost of production. If a plant manager is trying to decide whether it is profitable to produce power, he will compare the incremental revenue he will get from running the plant to the incremental cost of running the plant.[26] Until summer 2000 in California, suggestions that a $250 price cap would deter production were extremely difficult to credit. The incremental running costs of all plants in the state were well below this level. During the summer, the additional cost of air pollution permits in the south coast may have pushed the incremental cost for some plants in that area above the cap, and thus deterred them from producing. The problem became very salient in November and December of 2000, when a spike in the price of natural gas – rising from $4-$6 per million BTU to over $30 – put the incremental cost of nearly all gas plants above

[26] Amortization of the plant fixed costs will not figure into the calculation, but the need for more frequent maintenance as the plant operates for more hours will.

the price cap. These situations point out that price caps can in fact deter production. It also shows that price caps should be set in a way that takes into account variations in the cost of production. A single rigid price cap that is not indexed to costs of production will either have to be set so high that it has little effect or it will occasionally *cause* shortages and disruption in the market.

Price caps, however, can also deter the exercise of market power. A standard example of price caps demonstrates that a price cap set above the competitive price, but below the price that results without the cap, will lower prices and increase aggregate output from the firms in the market.[27] The intuition is that without price caps, firms with market power have an incentive to restrict output in order to increase price to their profit-maximizing level. If the price cap is imposed, then once the price reaches the cap, firms do not have an incentive to restrict output further in order to drive the price higher. Thus, the appropriate level for price caps trades off the risk of setting them too low and deterring production with the risk of setting them too high and permitting the exercise of excessive market power.

Even if price caps do not deter production, it has been argued that they can deter sale of that output in one area if the seller can receive higher prices in nearby areas that do not have a price cap. In all of the U.S. wholesale electricity markets, this has been a concern. However, if the price cap is set above the competitive price in an area that is a significant part of the regional grid, it is extremely unlikely that it will cause a shortage of power in the area that imposes the price cap. The basic reason is that at the *competitive* regional price, the quantity demanded can be provided by sellers without any generator selling at a price below its marginal cost. Thus, so long as the price is set higher than the regional competitive price, firms will have an incentive to provide all the power demanded across the grid without any firm selling at below its production cost.[28] Thus, a regional

[27] See Carlton & Perloff (1994), pp. 864-870, for an example.

[28] Consider a region with two markets, A and B, and a large-capacity transmission line between them. Assume that the competitive market price in this region would be $180, but a group of firms with market power are able to push the price up to $300. Assume also that marketers can take advantage of any profitable arbitrage opportunity between the two markets. Now impose a price cap of $250 in market A. Power then shifts to market B until the price in B is pushed down to $250. No more power would be sold in B – if it were, it would push the price in B below $250 – until all demand at $250 is satisfied in A. The question then is whether all demand at $250 will be satisfied. However, there is no reason for firms, even ones with market power, not to fulfill all demand at $250 because they can do so at a marginal cost below $250 (which follows immediately from the fact that the

price cap in the western U.S. is unlikely to do any more to cap prices than a price cap in California.

The long-run impact of price caps is easier to analyze conceptually, but more difficult to study empirically. A price cap will deter investment in new capacity if it is set, or investors believe it will be set, at a level that does not allow a return on investment that exceeds the investor's cost of capital. The data available on costs of building a power plant are necessarily rougher than the data on variable costs of production, because the costs of building a power plant are subject to many idiosyncratic factors related to location, siting restrictions, and other attributes. Furthermore, the beliefs of investors play a critical role, because the return is calculated over the life of the plant. Thus, just as under cost-of-service regulation, uncertainty about future regulatory intervention is likely to deter investment. It is for this reason primarily that price caps should be used with great caution. In a fully restructured electricity market with price-responsive demand and long-term contracts, price caps should exist only as a backstop measure, in case a failure occurs in the market mechanisms and the normal functioning of economic supply and demand is disabled.[29]

In discussing price caps thus far, I have assumed that a price cap once announced is credible and is never breached. That has not been the case in California, where the ISO has frequently violated the cap, both during the summer when the competitive price was probably below the cap nearly all of the time and in November and December when the competitive price almost certainly exceeded the cap much of the time. In the latter situation, violation of the cap was the only reasonable action since it clearly made more sense for many generators to shut down than to sell power at $250/MWh.

During the summer, however, the breaches of the cap made it very difficult to convince sellers that their efforts to raise the price above the cap through exercise of market power

competitive price is less than $250) and additional production to fulfill demand in A does not drive down the price they receive for their sales in B, which is already at $250. The only case in which this will not hold is if market A is so small relative to market B that all firms are better off ceasing sales in market A and maintaining the price above $250 in market B. (If the transmission line is not of sufficient capacity to handle all attempts to ship power out of the area with the price cap, then the cap is even less likely to result in shortages in the price cap area.)

[29] If, for instance, the dissemination of real-time price information broke down on a hot summer after (due, say, to an internet failure), the price cap would prevent prices from rising to levels that would not occur in a functioning market.

would not be successful. The discussion above relies critically on the ability of the buyer to stick to the cap. Otherwise, the announcement of the cap creates a game of "chicken" between sellers and buyers. In the case of California, the ISO's unwillingness to curtail demand, and inability to elicit demand-side response with real-time retail prices, put it in a very weak position in these showdowns. In California, the ISO breached the cap during summer 2000 by making "out-of-market" purchases at higher prices. Until recently, these purchases were restricted to sellers outside of the state. This set up an obvious strategy by generators to sell power out of the state and then resell it back in to the state at above the cap. A regional price cap – *if* it were credible and were set above the competitive price – could be quite valuable in deterring this behavior.

VI. Conclusion

California is currently trying to respond to extreme disruption in its electricity market. In part, the problems are attributable to factors that would have had adverse impacts even under the old regulatory regime: a spike in the price of natural gas, which has been much worse in the west than in the rest of the U.S.; environmental restrictions on generation in some areas; and tremendous economic growth in the state, and in other parts of the western grid, that has put real strain on the existing generation resources.

The restructuring of California's market, however, has greatly exacerbated the problems. The fact that commodity markets clear at a uniform price has meant that the mismatch between supply and demand, and the accompanying seller market power, has driven up the price for *all* power that the utilities must buy from other companies.[30] The inability of the utilities to buy much of that power through long-term contracts has undermined their ability to hedge against high short-term prices. The absence of restructuring on the demand side of the market, where virtually all consumption is still billed at a constant price regardless of the wholesale cost of power at the time it is supplied, has exacerbated the supply/demand mismatch and has increased the ability of sellers to exercise market

---

[30] I specify *other* companies here because, notwithstanding the recent order from the Federal Energy Regulatory Commission, a utility selling power into an electricity market and buying it back at the same price is not driving up its costs of providing power. When a utility sells power into the market and buys back the same amount of power from the same market at the same price, its *net* cost of that power is still just its production cost. The FERC's order that California utilities retain their power for their own use is not going to significantly change their costs of providing power to customers.

power.

Still, the movement towards restructuring of electricity markets was born from a history of well-supported dissatisfaction with outcomes under cost-of-service regulation. And it was recognized by many that the major benefits of restructuring would not occur until well after the time that the costs became apparent. Real-time retail pricing and long-term contracting will help to control the soaring wholesale prices receently seen in California, and will buy time to address other important structural problems that need to be solved to create a stable, well-functioning market. These problems include creating a workable structure for retail competition, determining the most efficient way to set locational prices and transmission charges, implementing a coherent framework for investing in new transmission capacity, and optimizing the ISO's procurement of ancillary services.

Those states and countries that have not yet started down this road would be wise to wait to learn more from the experiments that are now occuring in California, New York, Pennsylvania, New England, England and Wales, Norway, and Australia, as well as other locations. The difficulties with the outcomes so far, however, should not be interpreted as a failure of restructuring, but as part of the lurching process toward an electric power industry that is still likely to serve consumers better than the approaches of the past.

References

Allaz, B. and J.L. Vila, “Cournot Competition, Forward Markets, and Efficiency,” *Journal of Economic Theory*, **59**(February 1993), pp. 1-16.

Borenstein, Severin, “Understanding Competitive Pricing and Market Power in Wholesale Electricity Markets,” *Electricity Journal*, July 2000.

Borenstein, Severin and James Bushnell, “An Empirical Analysis of Market Power in a Deregulated California Electricity Market,” *Journal of Industrial Economics*, **47**(September 1999).

Borenstein, Severin, James Bushnell and Frank Wolak, “Diagnosing Market Power in California’s Deregulated Wholesale Electricity Market,” POWER Working Paper PWP-064, University of California Energy Institute, August 2000.

Carlton, Dennis W. and Jeffrey M. Perloff, *Modern Industrial Organization*, 2nd edition, New York: Harper Collins, 1994.

Joskow, Paul L., “Deregulation and Regulatory Refore in the U.S. Electric Power Sector,” in Peltzman and Winston, eds., *Deregulation of Network Industries: What’s Next?*, Washington, DC: AEI-Brookings Joint Center for Regulatory Studies, 2000.

Joskow, Paul L. and Edward Kahn, “A Quantitative Analysis of Pricing Behavior in California’s Wholesale Electricity Market During Summer 2000,” mimeo, November 2000.

Wolfram, Catherine D. (1999), “Measuring Duopoly Power in the British Electricity Spot Market,” *American Economic Review*, **89**(4), pp.805-826.

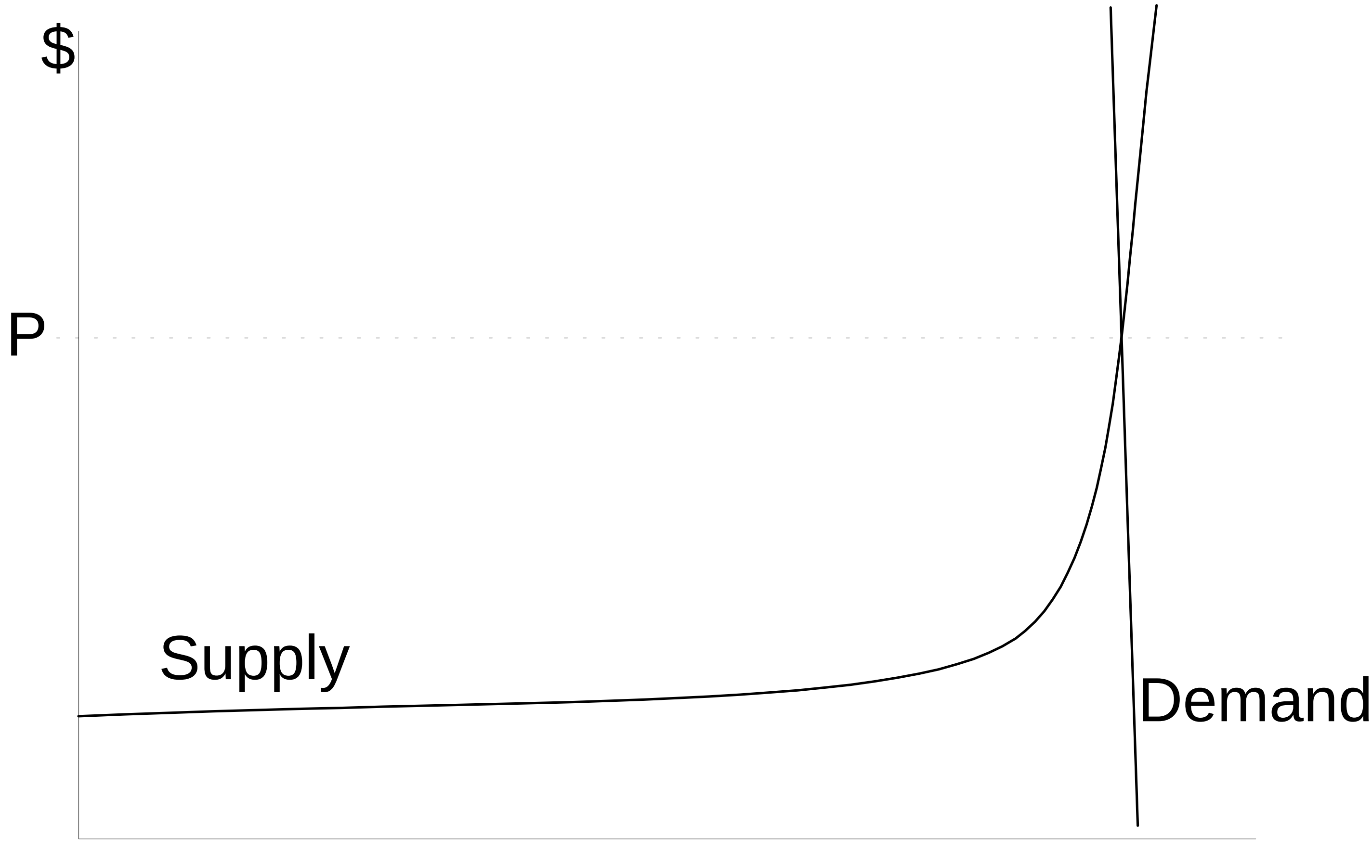


FIGURE 1

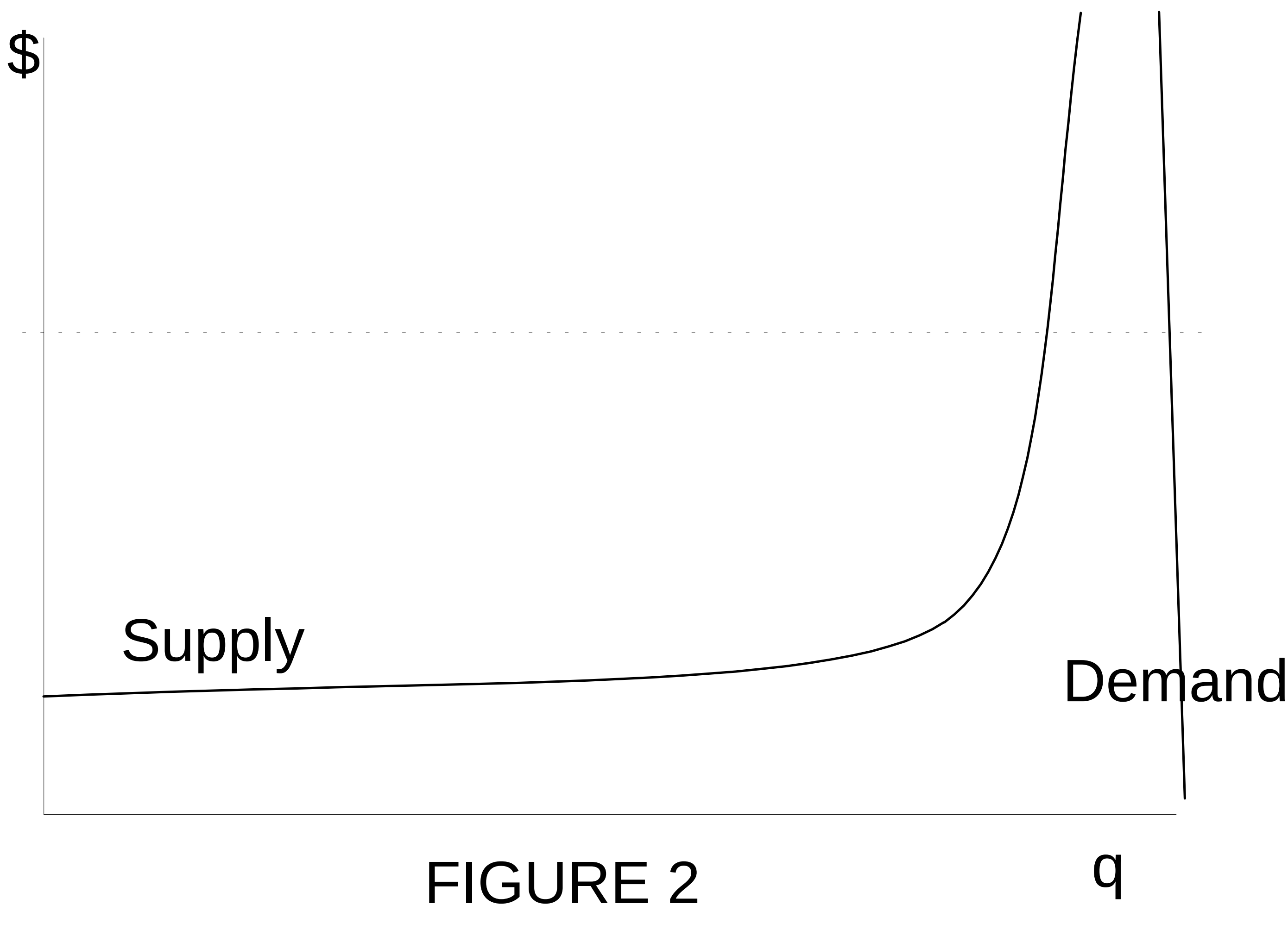


FIGURE 2

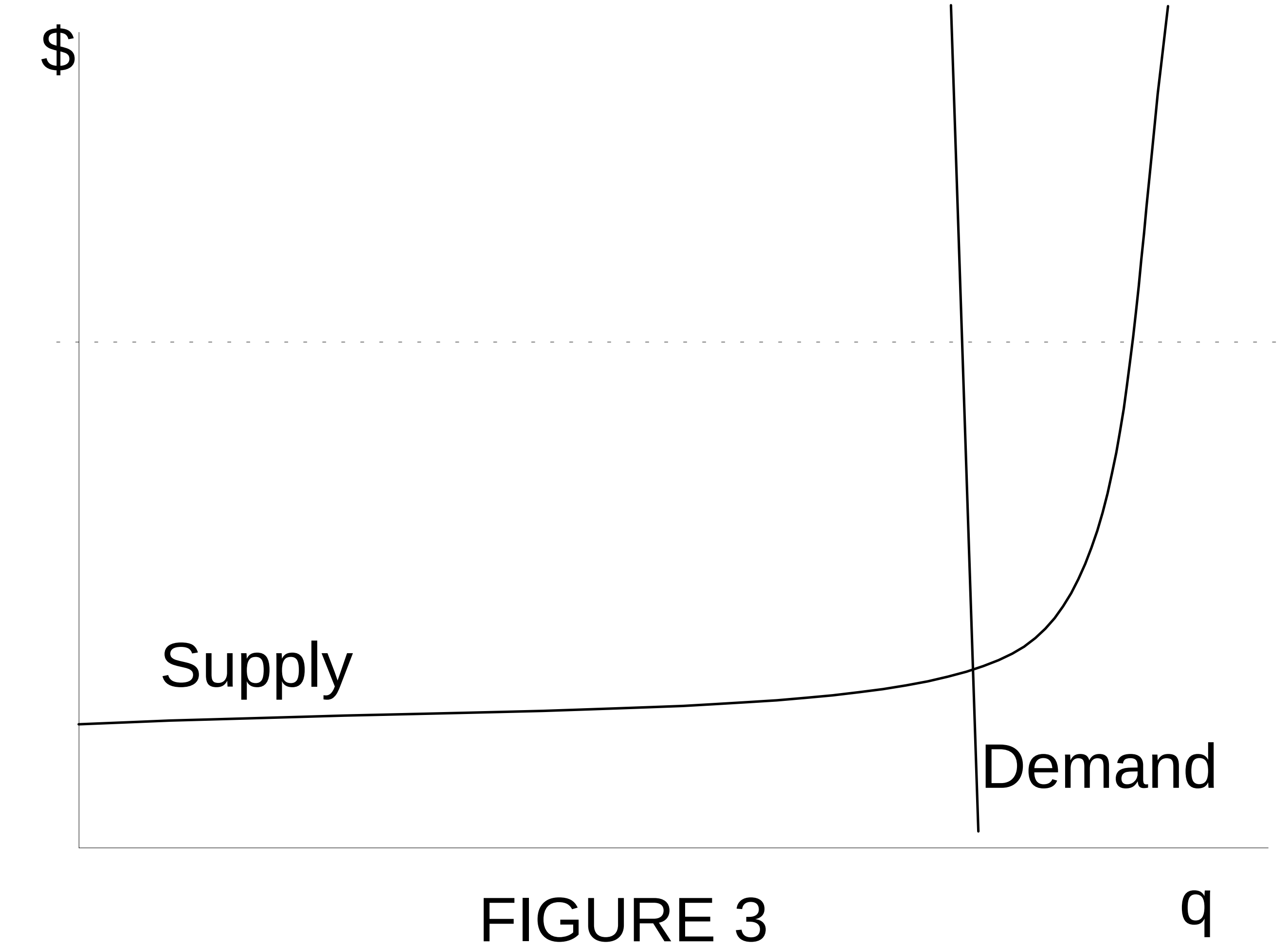


FIGURE 3

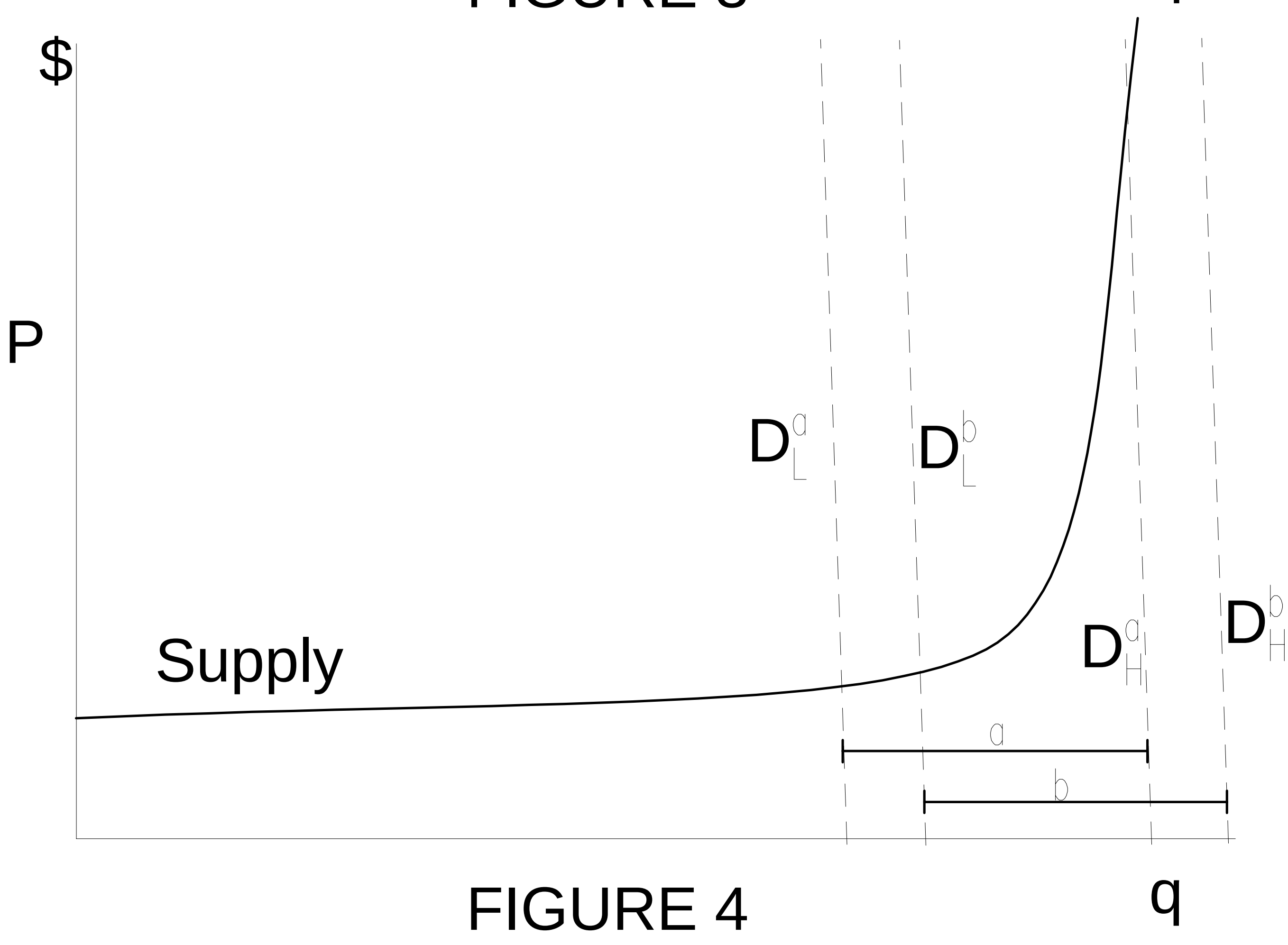
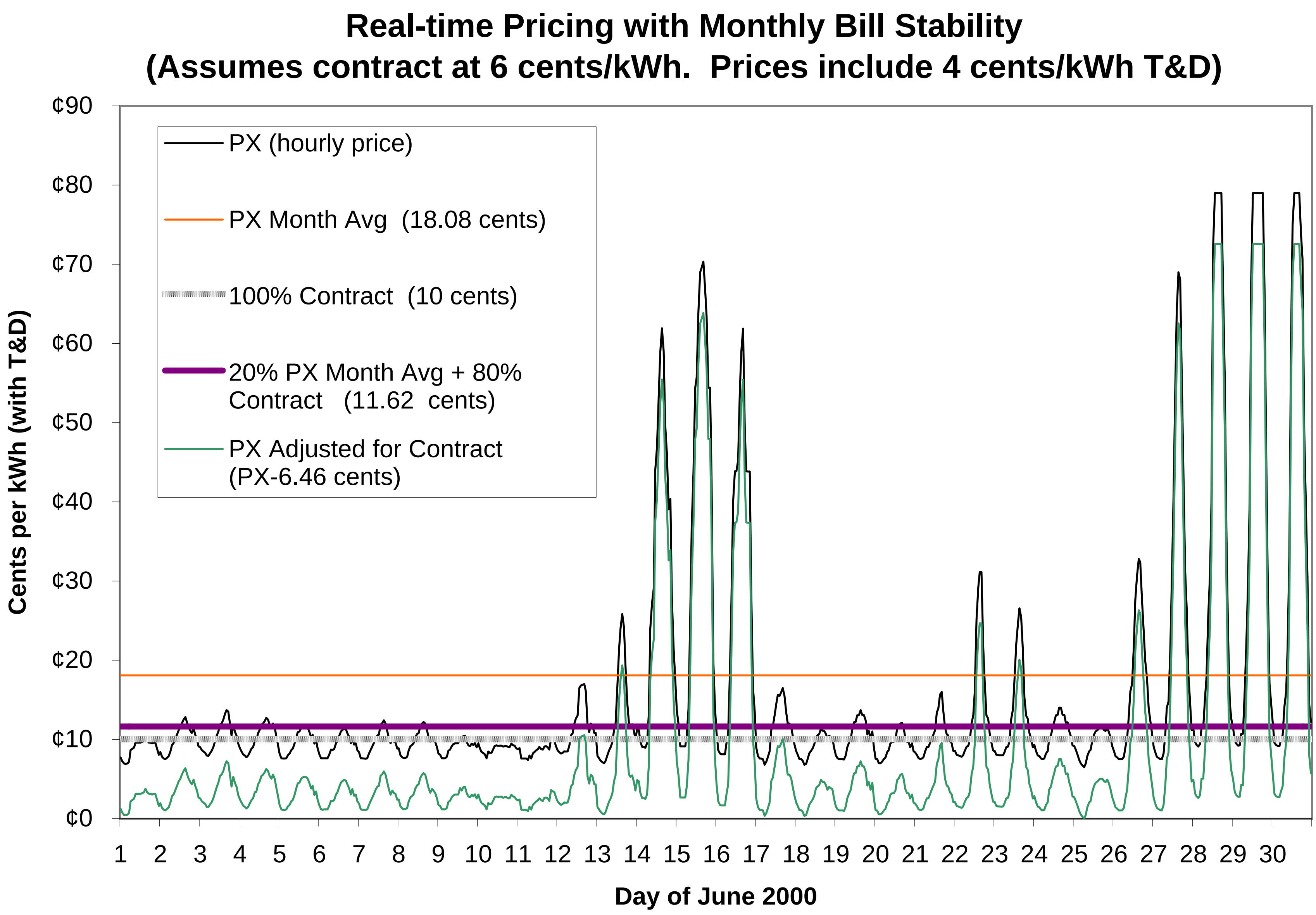


FIGURE 4



Real-time Pricing with Monthly Bill Stability
(Assumes contract at 6 cents/kWh. Prices include 4 cents/kWh T&D)
Cents per kWh (with T&D)
¢0
¢10
¢20
¢30
¢40
¢50
¢60
¢70
¢80
¢90
PX (hourly price)
PX Month Avg (18.08 cents)
100% Contract (10 cents)
20% PX Month Avg + 80% Contract (11.62 cents)
PX Adjusted for Contract (PX-6.46 cents)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Day of June 2000

FIGURE 5