Regulating Competition in Wholesale Electricity Supply

by
Frank A. Wolak
Department of Economics
Stanford University
Stanford, CA 94305-6072
wolak@zia.stanford.edu

First Draft: September 5, 2005
Revised: November 12, 2007

Abstract

The experience of the past ten years suggests that the potential benefits from electricity industry restructuring are small relative to those that can be achieved from introducing competition into other network industries such as telecommunications and airlines. In addition, the probability of a costly market failure in the electricity supply industry, often due to the exercise of unilateral market power, appears to be significantly higher than in other network industries. A major theme of this chapter is that electricity industry re-structuring is an evolving process that requires market designers to choose between an imperfectly competitive market and an imperfect regulatory process to provide incentives for least-cost supply at various of stages of the production process. The fundamental goal of the market design process in the wholesale market regime is to limit the ability of suppliers to exercise unilateral market power either explicitly through market price-setting mechanisms or implicitly through the regulatory price-setting process. There are a number ways the regulator can limit the ability of suppliers to exercise unilateral market power–namely, (1) alter the market structure, (2) change market rules, (3) impose penalties and sanctions on market participants for their behavior, and (4) even explicitly set the prices that market participants receive for their production. This chapter provides a theoretical framework for understanding how to make these choices in order to design a wholesale market that benefits consumers relative to the former vertically-integrated utility regime. The paper uses this framework to understand the causes of the disappointing experience with wholesale electricity restructuring in the US. This discussion points to a number of ways to increase the likelihood that restructuring in the US will ultimately benefit consumers.

1. Introduction

The technology of electricity production, transmission and distribution together with the history of pricing to final consumers make designing a competitive wholesale electricity market extremely challenging. There has been a number of highly visible wholesale market meltdowns, most notably the California electricity crisis during the period June 2000 to June 2001 and the sustained period of exceptionally high wholesale prices in New Zealand during June to September of both 2001 and 2003. Even wholesale markets generally acknowledged to have ultimately benefitted consumers relative to the former vertically-integrated monopoly regime in countries such as the United Kingdom and Australia have experienced substantial problems with the exercise of unilateral market power by large suppliers.

The experience of the past ten years suggests that, although there are opportunities for consumers to benefit from electricity industry re-structuring, these benefits have proved far more difficult to capture than those achieved from introducing competition into other network industries such as telecommunications and airlines. In addition, the probability of a costly market failure in the electricity supply industry, often due to the exercise of unilateral market power, appears to be significantly higher than in other formerly regulated industries. These facts motivate the three major questions addressed in this chapter. First, why has the experience with electricity structuring been so disappointing, particularly in the United States (US)? Second, what factors have led to success and limited the probability of costly market failures in other parts of the world? Third, how can these lessons be applied to improve wholesale market performance in the US and other industrialized countries?

An important theme of this chapter is that electricity industry re-structuring is an evolving process that continually requires market designers to choose on a going-forward basis between an imperfectly competitive market and an imperfect regulatory process to provide incentives for least-cost supply at all stages of the production process. As consequence, certain industry segments rely on market mechanisms to set prices and others rely on explicit regulatory price-setting processes. This choice depends on the technology available produce the good or service and the legal and economic constraints facing the industry.

Because the current technology for electricity transmission and local distribution overwhelmingly favors a single network for a given geographic area, a regulatory process is necessary to set the prices, or more the generally, the revenues transmission and distribution network

owners receive for providing these services. Paul Joskow's chapter in this volume, Joskow (2008), first presents the economic theory of incentive regulation–pricing mechanisms that provide strong incentives for transmission and distribution network owners to reduce costs and improve service quality and introduce new products and services in a cost-effective manner. He then provides a critical assessment of the available evidence on the performance of incentive regulation mechanisms for transmission and distribution networks.

The wholesale electricity segment of re-structured electricity supply industries primarily relies on market mechanisms to set prices, although the configuration of the transmission network and regulatory rules governing its use can exert a dramatic impact on the prices electricity suppliers are paid. In addition, the mechanism used to determine the location and magnitude of expansions to the transmission network has an enormous impact on the scale and location of new generation investments. Because a restructured electricity supply industry requires explicit regulation of certain segments and the regulatory mechanisms implemented significantly impact market outcomes, the entity managing the restructuring process must continually balance the need to foster fierce competition in those segments of the industry where market mechanisms are used to set prices against the need to intervene to set prices and control firm behavior in the monopoly segments of the industry. Maintaining this delicate balance requires a much more sophisticated regulatory process relative to the one that existed under the former vertically-integrated utility regime.

This chapter first describes the history of the electricity supply industry in the US and the motivation for the vertically-integrated monopoly industry structure and regulatory process that existed until wholesale markets were introduced in the late 1990s. This is followed by a description of the important features of the technology of supplying electricity to final consumers that any wholesale market design must take into account. These technological aspects of electricity production and delivery and the political constraints on how the industry operates make wholesale electricity markets extremely susceptible to the exercise of unilateral market power. This is the primary reason why continued regulatory oversight of the electricity supply industry is necessary and is a major motivation for the historic vertically-integrated industry structure.

To provide historical context for the electricity industry re-structuring process in the US, I then summarize the regulatory structure governing the electricity supply industry. This includes a description of the perceived regulatory failures than led to electricity industry re-structuring and a

description of the legal and regulatory structure governing the wholesale market regime in the US. In the vertically-integrated monopoly regime, the major challenge was providing incentives for the firms to produce in a least-cost manner and set prices that only recovered incurred production costs. Informational asymmetries about the production process or structure of demand between the vertically-integrated utility and the regulator made it impossible for the regulator to determine the least-cost mode of supplying retail customers.

In the wholesale market regime, the major challenge is putting in place a set of market rules that provides strong incentives for least-cost production by all suppliers and limits the ability of these suppliers to impact the market price through their unilateral actions. Different from the vertically-integrated regime, suppliers set market prices through their own unilateral actions that can result in market prices which deviate substantially from those necessary to recover production costs. To emphasize the difficulty in accomplishing this goal, this chapter describes the technological aspects of electricity production and delivery and the political constraints on how the industry operates which make wholesale electricity markets extremely susceptible to the exercise of unilateral market power.

I then introduce the generic wholesale market design problem as generalization of a multi-level principal-agent problem. There are two major dimensions to the market design problem. The first is public versus private ownership. The second is market versus explicit regulation to output prices. The impact of these choices on the principal-agent relationships between the firm and its owners and the firm and the regulatory body are discussed.

I then turn to the market design challenge in the wholesale market regime with privately-owned firms–limiting the ability and incentive of suppliers to exercise unilateral market power in the short-term wholesale market. There are a number ways the regulator can limit the ability of suppliers to exercise unilateral market power. To organize this discussion, I introduce the concept of a residual demand curve–the demand curve facing an individual suppliers after the responses of its competitors have been taken into account. I show that controlling the ability and incentive of suppliers to exercise unilateral market is equivalent to making the residual the supplier faces as price elastic as possible. I outline four things that the market designer can do increase the elasticity of the residual demand a supplier faces. I also argue that virtually all wholesale market meltdowns and

shortcomings of existing market designs can be traced to a failure to address adequately of one these dimensions of the market design process.

The final aspect of the market design process is the formation of an effective and credible regulatory oversight process for the industry. I argues that regulatory oversight of the wholesale market regime presents many extremely difficult challenges not faced by regulators in the former vertically integrated regime. A far more sophisticated regulatory process than the one that existed in the vertically-integrated regime is necessary for the wholesale marker regime. The regulator must engage in a process of continuous feedback and improvement in the market rules, which implies access to more information and more sophisticated use of the information provided.

The next section provides examples from wholesale markets in industrialized and developing countries of common market design flaws. These include excessive focus by the regulatory process on spot market design, inadequate divestiture of generation capacity by the incumbent firms, lack of an effective local market power mitigation mechanism, price caps and bid caps on short-term markets, and an inadequate retail market infrastructure.

The paper closes with a discussion of the causes of the disappointing experience with wholesale electricity markets in the US. There are number of economic and political constraints on the electricity supply industry in the US that have hindered the development of wholesale electricity markets that benefit consumers relative to the former vertically-integrated regime. This discussion points out a number of ways to increase the likelihood that electricity industry restructuring in the US will ultimately benefits consumers.

2. History of Electricity Supply Industry and the Path to Re-Structuring

This section reviews the history of the electricity supply industry in the US. I first address the origins of the vertically-integrated, regulated-monopoly industry structure that existed throughout the US until very recently. I then turn to a description of the factors that led to the recent re-structuring of the electricity supply industries in many parts of the US. In order to provide the necessary technical background to understand my analysis of the challenges facing wholesale market regime that takes the majority of this chapter, I describe important features of the technology and electricity production and delivery. I then turn to a discussion of the regulatory structure governing the electricity supply industry in the US how it has and has not yet evolved to dealt with the wholesale market regime.

2.1. A Brief Industry History to the Present

The electricity supply industry is typically divided into four stages: (1) generation, (2) transmission, (3) distribution, and (4) retailing. Generation is the process of converting raw energy from oil, natural gas, coal, nuclear power, hydro power, and renewable sources into electrical energy. Transmission is the bulk transportation of electricity in high voltages to limit the losses between the point at which the energy is injected into the transmission network and the point it is withdrawn from the network. In general, higher transmission voltages imply less energy losses over the same distance. Distribution is the process of delivering electricity at low voltage from the transmission network to final consumers. Retailing is the act of purchasing wholesale electricity and selling it to final consumers.

Historically, electricity supply for a given geographic area was provided by the single vertically-integrated utility that produced virtually all of the electricity it ultimately delivered to consumers. This firm owned and operated generation assets, the transmission network, and local distribution network required to deliver electricity throughout its geographic service area. There is some debate surrounding the rationale underlying the origins of this industry structure.

The conventional view is there are economies of scale in the generation and transmission of electricity at the level of demand served by most electricity utilities and significant economies of scope between transmission and distribution and generation at the level of demand and size of the geographic region served by most vertically-integrated utilities. The standard argument is that these economies to scale and scope create a natural monopoly, where the minimum cost industry structure to serve all consumers in given geographic area is a vertically-integrated monopoly. Without regulatory oversight, a large vertically-integrated firm could set prices substantially in excess of the marginal cost of the last unit sold.

The prospect of the large vertically-integrated firm using these economies to scale in transmission and generation and economies to scope between transmission and distribution and generation to exercise significant unilateral market power implies the need for regulatory oversight to protect the public interest and set the prices the monopoly supplier is allowed to charge and the terms and conditions under which it can charge these prices. What is often called the public interest rationale for the vertically-integrated, regulated-monopoly industry structure states that explicit output price regulation is necessary to protect consumers from the unilateral market power that could

be exercised by the dominant firm in a given geographic area. Viscusi, Vernon and Harrington (2001, Chapter 11) provides an accessible discussion of this perspective on the vertically-integrated, regulated-monopoly industry structure.

Jarrell (1978) proposes an alternative rationale for an industry composed of privately-owned, vertically-integrated monopolies subject to state-level regulation using the positive theory of regulation developed by Stigler (1971) and Peltzman (1976) . He argues that this market structure arose from the early years of the industry when utilities were regulated by municipal governments through franchise agreements. Many large municipalities issued duplicate franchise agreements and allowed firms to compete for customers. Jarrell argues that state-level regulation arose because these firms found it too difficult to maintain their monopoly status by their own actions, and instead decided to subject themselves to state-level regulatory oversight in exchange for a government-sanctioned geographic monopoly. Jarrell demonstrates that the predictions of the traditional public interest rationale for state regulation–that prices and profits should decrease and output should increase in response to state-level regulation–are contradicted by his empirical work. He finds statistically significantly higher output prices and profit levels and lower output levels for utilities in states that adopted state-level regulation early relative to those in states that adopted state-level regulation later. At a minimum, Jarrell's work suggests that the logic underlying state-level regulation of vertically-integrated monopolies is more complex than the standard public interest rationale described above.

Until the re-structuring process began in the late 1990s, the vast majority of US consumers were served by privately-owned vertically-integrated monopolies, although there were number of municipally-owned, vertically-integrated utilities and an even larger number of customer-owned electricity cooperatives serving rural areas. As noted in Joskow (1974), customers served by privately-owned vertically-integrated, regulated utilities experienced continuously declining real retail electricity prices from the start of the industry until the mid-1970s. Not until the second half of the 1970s did real electricity prices begin to increase and the vertically-integrated, regulated-utility industry structure begin to show signs of stress.

Perceived failures in this industry structure, technical change in electricity production and transmission and distribution, and events in electricity supply industries outside of the US all contributed to the impetus to re-structure. Joskow (1989) provides an perspicacious discussion of

the history of the US electricity supply industry and events leading up to the perceived failure of this regulatory paradigm and the initial responses to it. He argues that particularly in regions of the countries with rapidly growing electricity demand during the late-1970s and early 1980s, new capacity investment decisions made by the vertically-integrated utilities ultimately turned out to be extremely costly to consumers. This led to a general dissatisfaction with the vertically-integrated regulated-monopoly paradigm.

Around this same time technical change allowed generation units to realize all available economies to scale at significantly lower levels of capacity. For example, Joskow (1987) presents empirical evidence that scale economies in electricity production at the generation unit level are exhausted at a unit size of about 500 megawatts (MW)[1]. More recent econometric work finds that the null hypothesis of constant returns to scale in the supply of electricity (the combination of generation, transmission and distribution) by United States (US) investor-owned utilities cannot be rejected (Lee, 1995), which implies that economies to scope between transmission and generation are exhausted for the geographic areas served by most vertically-integrated utilities in the US.

During this same period a number of countries around the world were beginning the process of privatization and re-structuring of their state-owned, vertically-integrated electricity supply industry. In the late 1980s, England and Wales were the first industrialized countries to embark on this process. A number of other industrialized countries quickly followed–specifically, Norway, Sweden, Australia, and New Zealand. These international reforms demonstrated the feasibility of wholesale electricity competition and provided models for the re-structuring process in the US.

All of these factors combined to provide significant inertia in favor of the formation of formal wholesale electricity markets in the US. Joskow and Schmalensee (1983) provide a detailed analysis of the viability of wholesale competition in electricity as of the beginning of the 1980s.

2.2. Key Features of Technology of Electricity Production and Delivery

This section describes the basic features of electricity production, delivery, and demand. I first introduce the basic cost structure for electricity generation units. I then discuss how the form of a generation unit's cost function determines when it should operate in order to meet the pattern

[1]Typically there are multiple generation units at a single plant location. For example, at 1600 MW coal-fired plant may be composed of four 400 MW generation units at that site.

of hourly system demand throughout the year at least cost. The validity of this logic is demonstrated with examples of the actual average daily pattern of output of specific kinds generation units. I then explain the basic physics governing flows in electricity transmission networks, which considerably complicates the process of finding output rates for generation units to meet electricity demand at all locations in the transmission network.

Electricity production typically involves a significant up-front investment to construct a generation unit and a variable cost of producing electricity once the unit is constructed. Fossil fuel generation units using the same input fuel can be differentiated by their heat rate, the rate at that they convert heat energy into electrical energy. Heat rates are typically expressed in terms British Thermal Units (BTUs) of heat energy necessary to produce one kilowatt-hour (KWh) of electricity. For example, a natural gas-fired steam turbine unit might have a heat rate 9,000 BTU/KWh, whereas a natural gas-fired combustion turbine generation unit might have a heat rate 14,000 BTU/KWh. Lower heat rate technologies are typically associated with higher up-front fixed costs. Higher heat rate units are also typically less expensive to turn on and off. To convert a heat rate into the variable fuel costs associated with producing electricity, it is necessary to multiply the heat rate by the $/BTU price of the input fuel. For example, if the price of natural gas is $7 per million BTU, this implies a variable fuel cost of $63/MWh for the unit with at 9,000 BTU/KWh heat rate and a variable fuel cost of $98/MWh for the unit with 14,000 BTU/KWh heat rate. Other variable cost factors are added to the variable fuel cost to arrive at the unit's variable cost of production.

This relationship between the fixed and variable costs of producing electricity implies a total cost function for producing electricity at the generation unit level of the form $C_i(q) = F_i + c_i q$, where F_i is the up-front fixed cost and c_i is the variable cost of production for unit i. In general, the total variable cost of producing electricity is nonlinear in the level of output.[2] Simplifying the general nonlinear variable cost function $vc_i(q)$ to the linear form $c_i q$ makes it more straightforward to understand when during the day and year a generation unit will operate. Suppose there are two generation units, with $F_1 > F_2$ and is $c_1 < c_2$, consistent with above logic that a lower variable cost of production comes at the expense of a higher fixed cost of production. In order for the total costs

[2]Wolak (2007a) estimates generation unit-level daily variable cost functions for units participating in a wholesale electricity market and finds strong evidence of economically significant nonlinearities both within and across periods of the day in the variable cost of producing electricity.

of operating unit 1 during the year to be less than the total cost of operating unit 2 during the year, unit 1 must produce more than q^*, where q^* solves the following equation in q:

$$F_1 + c_1 q = F_2 + c_2 q, \text{ which implies } q^* = (F_1 - F_2)/(c_2 - c_1).$$

At higher levels of annual output than q^*, total annual production costs are less for unit 1 than for unit 2. This is because the variable cost of producing output from unit 1 is less than the variable cost of producing output from unit 2 for all output levels above q^*. Conversely, total annual production costs are lower if unit 2 is used to produce all output levels below q^*. These facts are useful to understand the least cost mix of production from the available generation unit technologies needed to meet the annual distribution of half-hourly or hourly electricity demands.

The annual pattern of half-hourly or hourly electricity demands is usually represented as a load duration curve. Figure 1 plots the half-hourly load duration curve for the state of Victoria in Australia for three years: 2000, 2001 and 2002. The Victoria market operates on a half-hourly basis, so each point on an annual load duration curve gives the number of half-hours during the year on the horizontal axis that demand is greater than or equal to value on the vertical axis. For example, for 8,000 half-hours of the year in 2000, system demand is greater than or equal to 5,500 MW. For both 2001 and 2002, for 8,000 half-hours of the year demand is greater than or equal to 6,000 MW.

The load duration curve can be used to determines how the mix of available generation units should be used to meet this distribution of half-hourly demands at least cost. Generation units with the lowest variable costs will operate during all half-hours of the year. This is represented on the load duration curve by a rectangle with height equal to the average half-hourly output of the unit and length equal to the number of half-hours in the year. Rectangles of this form are added on top of one another from the lowest to the highest annual average cost of production until the rectangular portion of the load duration curve is filled–the level of system demand that is exceeded during all half hours of the year. Additional rectangles of increasingly smaller lengths of operation are stacked up from the lowest to highest annual average cost of providing the desired amount of annual energy until the load duration curve is covered by these rectangles. This process of filling the load duration curve implies that higher variable costs unit should be called upon less frequently that lower variable cost units.

This logic has implications for how the daily pattern of half-hourly demands are met. Figure 2 plots the annual average daily pattern of demand for Victoria for the same three years as Figure 1. A point on the curve for each year gives the annual average demand for electricity in MW for the half-hour period during the day given on the horizontal axis. For example, during half-hour period 20 of the year 2000, the annual average half-hourly load is 5,500 MW. This half-hourly pattern of load within the day and the process used to fill the load-duration curve described above imply different patterns of half-hourly output within the day for specific generation units depending on their cost structure. Figure 3 plots the average daily pattern of output from the Yallourn plant for 2000, 2001 and 2002. This plant is composed of four brown coal units that produce output at a variable cost of approximately 5 Australian dollars ($AU) per MWh. As discussed in Wolak (2007c), these units have the lowest variable cost in Australia and by the above logic of filling the load duration curve, they should operate at the same level during all hours of the day. As predicted by this logic, Figure 3 shows that for each of the three years, there is little difference in the average half-hourly output level across half hours of the day.

Figure 4 plots the average daily pattern of output from Valley Power plant for 2002. This plant came on line in November 2001 and is composed of six generation units totaling 300 MW. Each of these units has one of the highest variable costs in Victoria. By the logic of filling the load duration curve, these units should operate only in the highest demand periods of the day. Figure 2 shows that average half-hourly demand in Victoria is highest around period 30. The average half-hourly output of the Valley Power plant is highest in period 30 and slightly lower in the surrounding half-hours and declines to close to zero in the remaining half-hours of the day, which is consistent with the logic of filling the load duration curve.

A final aspect of the load duration curve has implications for the cost effectiveness of active demand-side participation in the wholesale market. Figure 5 plots the load duration curve for highest 500 half-hour periods for the same three years as Figure 1. This figure shows that the load duration curve for 2002 intersects the vertical axis at approximately 7,600 MW. At a value on the horizontal axis of 10 half-hours, the value of the curve falls to approximately 7,400 MW, which means that least 200 MW of generation capacity is required to operate at less than 10 half-hour periods of the year. If system demand could be reduced below 7,400 MW during these 10 half-hour periods, through demand response or energy efficiency programs, this would eliminate the need to

construct and operate a peaking generation facility such as the Valley Power plant. An extremely steep load duration curve near the vertical axis implies that a substantial amount of capacity is used a very small number of hours of the year, and that there would be significant saving in construction and operating costs by providing final consumers with incentives to reduce their demand during these hours.

Perhaps the most important feature of wholesale electricity markets is that the unilateral actions of generation unit owners to raise wholesale prices can result in a substantial divergence between the market-clearing price and variable cost of the highest cost unit operating during that half-hour period (the price that would arise if no supplier could exercise unilateral market power). Figure 6 plots the annual daily average of half-hourly prices for Victoria for 2000, 2001, and 2002. The extremely high annual average price during half-hour 30 for 2002 illustrates the extent to which there can a divergence between the variable cost of highest cost unit operating during a half-hour and the market-clearing price. As noted above, the variable cost of producing electricity from peaking units such as the Valley Power plant depends primarily on the price of natural gas. However, the price of natural gas in Victoria changed very little from 2000 to 2002, but the annual average price of electricity for half-hour period 30 and the surrounding half-hour periods for 2002 are substantially above the annual average prices for the same half-hour periods in 2000, and the annual average price for half-hour period 30 and the surrounding half-hours for 2000 are significantly above the annual average prices for the same half-hour periods in 2001. These differences in annual average half-hourly prices across the years demonstrate that competitive conditions and other factors besides the variable costs of the highest cost unit operating are major drivers of the level of average electricity prices in a wholesale market regime.

A final distinguishing feature of the electricity supply industry is the requirement to deliver electricity through a potentially congested looped transmission network. Electricity flows along the path of least resistence thorough the transmission network according to Kirchhoff's First and Second Laws rather than according to the desires of buyers and sellers of electricity.[3] To understand the operation of looped electricity networks, consider the three-node network in Figure 7. Assume that links AB, BC and AC have the same resistence and that there are no losses associated with

[3]See http://physics.about.com/od/electromagnetics/f/KirchhoffRule.htm for an accessible introduction to Kirchoff's Laws.

transmitting electricity in this network. Suppose a supplier located at node A injects 90 megawatts (MW) of energy for a customer at node B to consume. Kirchoff's Laws imply that 60 MW of the 90 MW will travel along the link AB and 30 MW will travel along the pair of links AC and BC because the total resistence along this indirect path from A to B is twice the resistence of the direct path from A to B.

How this property of a looped transmission network impacts wholesale market outcomes becomes clear when the capacities of transmission links are taken into account. Suppose that the capacity of link AB is 40 MW, and the capacities of links AC and BC are each 100 MW. Ignoring the physics of power flows, one might think that the capacity of the AC and BC links would allow injecting 90 MW at node A and withdrawing 90 MW at node B. Kirchoff's laws imply that the maximum amount of energy that can be injected at A and withdrawn at node B is 60 MW, because 40 MW will flow along AB and 20 MW will flow along the links AC and BC. The 40 MW capacity of link AB limits the amount that can injected at node A. For this configuration of the network, the only way to allow consumers at node B to withdraw 90 MW of energy would be to inject less energy at node A and more at node C, so that the total injected at A and C is equal to 90 MW. For example, injecting 30 MW at node A and 60 MW at node C would result in a flow of 40 MW on link AB and allow total withdrawals of 90 MW at node B.

2.3. Regulatory Transition from Vertically-Integrated Utility Regime

Regulatory oversight in the US is complicated by the fact that the federal government has jurisdiction over interstate commerce and state governments have jurisdiction over intrastate commerce. This logic implies that state governments have the authority to regulate retail electricity prices and intrastate wholesale electricity transactions, and the federal government has the authority to regulate interstate wholesale electricity transactions.

The physics of electricity flows in a looped transmission network does not allow a clear distinction between interstate and intrastate sales of electricity arbitrary. As described above, electricity flows according to the path of least resistance, which makes it extremely difficult, if not impossible, to determine precisely how much of the electricity consumed in one state was actually produced in another state if the two states are interconnected by a looped transmission network. This has led to a number of rules of thumb to determine whether a wholesale electricity transaction is subject to federal or state jurisdiction. Clearly, trades between parties located in different states

is subject to federal oversight. However, it also possible that a transaction between parties located in the same state is subject to federal oversight. One determinant of whether a transaction among parties located in the same state is classified as interstate and subject to federal oversight is the voltage of the transmission lines that the buyer and seller use to consummate the deal, because as discussed above, higher voltage lines usually deliver more electricity over longer distances.

The Federal Power Act of 1930 established the Federal Power Commission (which became the Federal Energy Regulatory Commission (FERC) in 1977) to regulate wholesale energy transactions using high-voltage transmission facilities. The Federal Power Act established standards for wholesale electricity prices that FERC must maintain. In particular, FERC is required to ensure that wholesale electricity prices are "just and reasonable." Prices that only recover the supplier's production costs, including a return to capital, meet the just and reasonable standard. Prices set by other means can also meet this standard because FERC is the ultimate authority on what constitutes a just and reasonable wholesale price. If FERC determines that wholesale electricity prices are not just and reasonable, then it also has considerable discretion to take actions to make these prices just and reasonable, and it must order refunds for any payments made by consumers at prices in excess of just and reasonable levels.

It is important to emphasize that these provisions of the Federal Power Act have not been repealed, despite the existence of bid-based wholesale electricity markets throughout the northeast US, parts of the Midwest and in California. As discussed below, the requirement that wholesale electricity prices satisfy the "just and reasonable" standard of the Federal Power Act is a major challenge to introducing wholesale competition in the US.

Under the vertically-integrated regime, state-level regulation of retail electricity rates effectively controls the price utilities pay for wholesale electric power. Utilities either own all of the generation units necessary to meet their retail load obligations or supplement their generation ownership with long-term contract commitments for energy sufficient to meet their retail load obligations. The implicit regulatory contract between the state regulator and the utilities within its jurisdiction is that in exchange for being allowed to charge a retail price set by the regulator that allows the utility the opportunity to recover all prudently incurred costs of building and operating their generation units and paying prudently incurred long-term energy contract costs, the utility has an obligation to serve all demand in its geographic service area at this regulated price. Although

these vertically integrated utilities sometimes make short-term electricity purchases from neighboring utilities, virtually all of their retail energy obligations are met either from long-term contracts or generation capacity owned and operated by the utility.

The vertically-integrated utility industry structure and state-level regulation of retail prices makes federal regulation of wholesale electricity transactions largely redundant. The state regulator does not allow utilities under its jurisdiction to enter into long-term contracts that it does not believe are in the interests of electricity consumers in the state. Therefore, under the vertically-integrated state-regulated monopoly industry structure, FERC's regulatory oversight of wholesale prices often amounts to no more than approving transactions deemed just and reasonable by a state regulator. This implicit state-level regulation of wholesale prices caused FERC to have very little experience regulating wholesale electricity transactions when the first formal wholesale markets began operation the US in the late 1990s.

Joskow (1989) notes that there are a number of flaws in the state-level regulatory process that created advocates for the introduction of formal wholesale markets. First, retail electricity prices are only adjusted periodically, at the request of the utility or state commission, and only after a lengthy and expensive administrative process. Because of the substantial time and expense of the review process, utilities and commissions typically wait until this time and expense can be justified by a large enough expected price change to justify this effort. Consequently, the utility's prices typically track the utility's costs very poorly. This regulatory lag between price changes and cost changes can introduce incentives for cost minimization on the part of the utility during periods when input prices increase. As Joskow (1974) describes in detail, nominal prices remained unchanged for a number of years during the 1950s and 1960s. This is primarily explained by both gains in productive efficiency and utilities exploiting economies of scale and scope in electricity supply during a period of stable input prices.

During the late 1970s and early 1980s when input fossil fuel costs rose dramatically in response to rapidly increasing world oil prices, many utilities filed to increase their prices a number of times in rapid succession. Joskow (1974) emphasizes that state regulators are extremely averse to nominal price increases and they have considerable discretion to determine what costs are prudently incurred and the utility is entitled to recover in the prices it is allowed to charge. Consequently, a rational response by the regulator to nominal input cost increases is to grant output

price increases lower than the utility requested. Disallowing cost recovery of some investments is one way to accomplish this. Joskow (1989) stresses that the "used and useful" regulatory standard is the basis for determining whether an investment is prudent. If an asset is used by the utility and useful to produce its output in a prudent manner, then this cost has been prudently incurred. Clearly there is some circularity to this argument, and that can allow regulators to disallow cost recovery for certain investments that seemed necessary at the time they were made but subsequently turned out not to be necessary to serve their customers.

Joskow (1989) states that as result of the enormous cost increases faced by utilities during the mid-1970s and early 1980s, a number generation investments at this time were subject to ex post prudence reviews by state public utilities commissions (PUCs), particularly when the forecasted enormous increases in fossil fuel prices used to justify investments in coal-fired and nuclear generation facilities failed to materialize. Increasing retail electricity rates enough to pay for these investments was politically unacceptable, particularly given the reduction in fossil fuel prices that occurred in the mid-1980s onward. The utility's shareholders had to cover the losses associated with these coal and nuclear generation unit investments that were deemed by the state PUC to be ex post imprudent. As a consequence, the utility's appetite for investing in large baseload generation facilities, even in regions with significant demand growth, was substantially reduced.

Joskow concludes his discussion of these events with the following statements.

> The experience of the 1970s and early 1980s has made it clear that existing industrial and administrative arrangements are politically incompatible with rapidly rising costs of supplying electricity and uncertainty about costs and demand. The inability of the system to deal satisfactorily with these economic shocks created a latent demand for better institutional arrangements to regulate the industry, in particular to regulate investments in and operation of generation faciliities." (Joskow, 1989, p. 162).

This experience began the process of re-structuring of the electricity supply industry in the US. Joskow (2000a) describes the transition from a limited amount of competition among cogeneration facilities and small scale generation facilities to sell wholesale energy to the vertically-integrated utility enabled by the Public Utilities Regulatory Policy Act (PURPA) of 1978 to the formation of formal bid-based wholesale markets, which first began operation in California in April of 1998.

Before closing this section, it is important to emphasize two key features of the regulatory process governing electricity supply in the US that will play a role in later discussions. First, for the reasons noted above, FERC historically had minor role in regulating wholesale electricity prices in the US and was largely unprepared for many of the challenges associated with regulating wholesale electricity markets. Joskow (1989) points out that over the decade of the 1980s "FERC staff has been increasingly willing to accept mutually satisfactory negotiated coordinated contracts between integrated utilities that are de facto unencumbered by the rigid cost accounting principles used to set retail rates." (p. 138). As described below, the fact that most of the generation capacity and the transmission and distribution assets used to serve the utility's customers were owned by the utility, combined with FERC's approach to regulating wholesale energy transactions, meant that state PUCs exerted almost complete control over retail electricity prices.

The advent of wholesale electricity markets with significant participation by pure merchant suppliers–those with no regulated retail load obligations–severely limited the ability of state PUCs to control retail prices. FERC's role in controlling wholesale and retail prices is increased by the extent to which the state-regulated load-serving entities no longer own generation assets and they purchase their wholesale energy needs from short-term markets. The California re-structuring process created a set of circumstances where FERC's role in regulating wholesale prices was far greater than in any of the wholesale markets in the eastern US. The major load-serving entities were required to sell virtually all of their fossil-fuel generation assets to merchant suppliers and the vast majority of wholesale energy purchases to serve their retail load obligations were made through short-term markets. Therefore, not by a conscious decision, the California re-structuring process resulted in California Public Utilities Commission (CPUC) giving up virtually all ability to control wholesale and retail prices in the state.

A second important feature of the regulatory process in the US is that the Federal Power Act still requires FERC to ensure that wholesale prices are just and reasonable, even if prices are set through a bilateral negotiation or through the operation of a bid-based wholesale electricity market. FERC recognizes that markets can set prices substantially in excess of just and reasonable levels, typically because suppliers are exercising unilateral market power. FERC has also established that just and reasonable prices are set through market mechanisms where no supplier exercises unilateral market power. Wolak (2003b and 2003d) discusses the details of how FERC uses this logic to

determine whether to allow a supplier to sell at market-determined prices, rather than cost-of-service prices. If a supplier can demonstrate that it has no ability to exercise unilateral market power or there are mechanisms in place that mitigate its ability to exercise unilateral market power, the supplier can sell at market-determined prices. FERC uses an elaborate market structure-based procedure to make this assessment. Wolak (2003b) points out a number of flaws in this procedure. Bushnell (2005) discusses an alternative approach that makes use of oligopoly models and demonstrates its usefulness with an application to the California electricity market.

3. Why Wholesale Electricity Markets Require Industry-Level Regulatory Oversight

This section describes the characteristics of the technology of electricity supply and the political and economic constraints facing the industry that make it extremely difficult to design wholesale electricity markets that consistently achieve competitive outcomes–market prices close to those that would be predicted by price-taking behavior by market participants. The extreme susceptibility of wholesale electricity markets to the exercise of unilateral market power and the massive wealth transfers from consumers to producers that can occur in a very short period of time as result, make regulatory oversight beyond that provided by antitrust law essential. The remainder of this section contrasts the major challenges facing the regulatory process in the wholesale market regime relative to the vertically-integrated, regulated monopoly regime.

3.1. Why Electricity is Different from other Products

It is difficult to conceive of an industry more susceptible to the exercise of unilateral market power than electricity. It possesses virtually all of the product characteristics that enhance the ability of suppliers to exercise unilateral market power.

Supply must equal demand at every instant in time and each location of the network. If this does not occur then the transmission network can become unstable and brownouts and blackouts can occur such as the one that occurred in the eastern US and Canada on August 13, 2003. It is very costly to store electricity. Constructing significant storage facilities typically requires substantial up-front costs and storing 1 MWh of energy requires consuming more than 1 MWh. Production of electricity is subject to extreme capacity constraints in the sense that it is impossible to get more than a pre-specified amount of energy from a generation unit in an hour.

As noted in Section 2.2, delivery of the product consumed must take place through a potentially congested, looped transmission network. If a supplier owns a portfolio of generation

units connected at different locations in the transmission network, how these units are operated can congest the transmission path into given geographic area and thereby limit the number of suppliers able to compete with those located on the other side of the congested interface. The example presented in Figure 7 with the capacity of link AB equal to 40 MW and the capacities of links AC and BC each equal to 100 MW illustrates this point. If all of a supplier's generation units are located at node A and all load is at node B, the firm at node A can supply at most 60 MW of energy to final consumers. If demand at node A is greater than 60 MW, then the additional energy must come from a supplier at node B. For example, if the demand at node B is 100 MW, because the capacity of the transmission link AB is 40 MW, the supplier at node B is monopolist facing a residual demand of 40 MW, if the supplier at node A is providing 60 MW,

Historically, how electricity has been priced to final consumers makes the wholesale demand extremely inelastic, if not perfectly inelastic, with respect to the hourly wholesale price. In the US, customers are typically charged a single fixed price for each kilowatt-hour (KWh) they consume during the month regardless of the value of the wholesale price when each KWh is consumed. Paying according to fixed-retail price implies that these customers have hourly demands with a zero price elasticity with respect to the hourly wholesale price. The primary reason for this approach retail pricing is that most residential meters are only capable of recording the total amount of KWh consumed between consecutive meter readings, which typically occur at monthly intervals. Consequently, a significant economic barrier to setting retail electricity prices that reflect wholesale market conditions is the availability of a meter on the customer's premise that records hourly consumption for each hour of the month.

There is growing empirical evidence that all classes of customers can respond to short-term wholesale price signals if they have the metering technology to do so. Patrick and Wolak (1997) estimate the price-responsiveness of large industrial and commercial customers in the United Kingdom to half-hourly wholesale prices and find significant differences in the average half-hourly demand elasticities across types of customers and half-hours of the day. Wolak (2006) estimates the price-responsiveness of residential customers to a form of real-time pricing that shares the risk of responding to hourly prices between the retailer and the final customer. The California Statewide Pricing Pilot (SPP) selected samples of residential, commercial, and industrial customers and subjected them to various forms of real-time pricing plans in order to estimate their price

responsiveness. Charles River Associates (2004) analyzed the results of the SPP experiments and found precisely estimated price responses for all three types of customers.

Although all of these studies find statistically significant demand reductions in response various forms of short-term price signals, none are able to assess the long-run impacts of requiring customers to manage short-time wholesale price risk. Wolak (2007a) describes the increasing range of technologies available to increase the responsiveness of a customer to short-term price signals. However, customers have little incentive to adopt these technologies unless state regulators are willing to install hourly meters and require customers to manage short-term price risk.

For the reasons discussed in Section 7, the vast majority of utilities that have managed to install hourly meters on the premises of some of their customers find it extremely difficult to convince state PUCs to require these customers to pay retail prices that vary with wholesale market conditions. Wolak (2007a) offers an explanation for this regulatory outcome and suggests a process for overcoming the economic and political constraints on more active demand-side participation in short-term wholesale electricity markets.

A final factor enhancing the ability of suppliers to exercise unilateral market power is that the potential to realize economies of scale in electricity production historically favored large generation facilities, and in most wholesale markets the vast majority of these facilities are owned by a relatively small number of firms. This generation capacity ownership also tends to be concentrated in small geographic areas within these regional wholesale markets, which increases the potential for the exercise of unilateral market power in smaller geographic areas.

All of the above factors also make wholesale electricity markets substantially less competitive the shorter the time lag is between the date the sale is negotiated and the date delivery of the electricity occurs. In general, the longer is the time lag between the agreement to sell and the actual delivery of the electricity, the larger the number of suppliers that are able to compete to provide that electricity. For example, if the time horizon between sale and delivery is more than two years, then in virtually all parts of the US new entrants can compete with existing firms to provide the desired energy. As the time horizon between sale and delivery shortens, more potential suppliers are excluded from providing this energy. For example, if the time lag between sale and delivery is only one month, then it hard to imagine that a new entrant could compete to provide this electricity.

It is virtually impossible to site, install and begin operating even a very small new generation unit in one month.

Although it is hard to argue that there is a strictly monotone relationship between the time horizon to delivery and the competitiveness of the forward energy market, the least competitive market is clearly the real-time energy market because so few suppliers are able to compete to provide the necessary energy. Only suppliers operating their units in real-time with unloaded capacity or quick-start combustion turbines at locations in the transmission network that can actually supply the energy needed are able to compete to provide it.[4]

For this reason, real-time prices are typically far more volatile than day-ahead prices, which are far more volatile than month-ahead or year-ahead prices. It is easy to imagine that an electricity retailer would be willing to pay $1,000/MWh for 10 MWh in the real-time market, or even $5,000/MWh, if that meant keeping the lights on for its retail customers. However, it is unlikely that this same load-serving entity would pay much above the long-run average cost of production for this same 10 MWh electricity to be delivered two-years in the future, because there are many new entrants willing to sell this energy at close to the long-run average cost of production.

This logic illustrates that system-wide market power in wholesale electricity markets is a relatively short-lived phenomenon if the barriers to new entry are sufficiently low. If system conditions arise that allow existing suppliers to exercise unilateral market power in the short-term market, they are also able to do so to varying degrees in the forward market at time horizons to delivery up to the time it takes for significant new entry to occur. In most wholesale electricity markets, this time horizon is between 18 months to 2 years, meaning that if system conditions arise that create opportunities for suppliers to exercise unilateral market power in the short-term energy market, unless these system conditions change or are expected to change in the near future, then suppliers can also exercise unilateral market power in the forward market for deliveries up to 18 months to 2 years into the future.[5] Although these opportunities to exercise system-wide market

---

[4]A generation unit has unloaded capacity if its instantaneous output is less than the unit's maximum instantaneous rate of output. For example, a unit with a 500 MW maximum instantaneous rate of output (capacity) operating at 400 MW has 100 MW of unloaded capacity.

[5]Wolak (2003b) documents this phenomenon for the case of the California electricity market during the Winter of 2001. Energy purchased at that time for delivery during the Summer of 2003 sold for approximately $50/MWh, whereas energy to be delivered during the Summer of 2001 sold for approximately $300/MWh and the Summer 2002 for approximately

power are relatively short-lived, the experience of a number of wholesale electricity markets has demonstrated that suppliers with unilateral market power are able to raise market prices substantially during this time period, which can lead to enormous wealth transfers from electricity consumers to producers, even for periods as short as three months.

Electricity suppliers possess differing degrees of systemwide and local market power. Systemwide market power arises from the capacity constraints in the production and the inelasticity of the aggregate wholesale demand for electricity, ignoring the impact of the transmission network. Local market power is the direct result of the fact that all electricity must be sold through a transmission network with finite carrying capacity. The geographic distribution of generation ownership and demand interact with the structure of the transmission network to create circumstances when a small number of suppliers or even one supplier is the only one able to meet an energy need at a given location in the transmission network.

The distinction between system-wide and local market power is often blurred by the choice of the relevant market. If electricity did not need to be delivered through a potentially congested transmission network subject to line losses, then it is difficult to imagine that any supplier could possess substantial system-wide market power if the relevant geographic market was the entire US. There are a large number of electricity suppliers in the US, none of which controls a significant fraction of the total installed capacity in the US. Consequently, the market power that an electricity supplier possesses fundamentally depends of the size of the geographic market it competes in, which depends on the characteristics of the transmission network and location of final demand.

Borenstein, Bushnell and Stoft (2000) demonstrate this point in the context of a two-node model of quantity-setting competition between suppliers at each node potentially serving demand at both nodes. They find that small increases in the capacity of the transmission line between the two locations can substantially increase the competitiveness of market outcomes at the two locations. One implication of the Borenstein, Bushnell and Stoft (2000) results is that a supplier possesses local market power regardless of the congestion management protocols used by the wholesale market. In single-price markets, zonal-pricing markets, and nodal-pricing markets, local market power arises because the existing transmission network does not provide the supplier with

---

$150/MWh.

sufficient competition to discipline its bidding behavior into the wholesale market.[6] This is particularly the case in the US, where the rate of investment in the transmission network has persistently lagged behind the rate of investment in new generation capacity over the past 25 years. Hirst (2004) documents this decline in the rate of investment in transmission capacity.

Most of the existing transmission networks in the US were designed to support a vertically-integrated utility regime that no longer exists. Particularly around large population centers and in geographically remote areas, the vertically-integrated utility used a mix of local generation units and transmission capacity to meet the annual demand for electricity in the region. Typically, the utility supplied the region's baseload energy needs from distant inexpensive units using high-voltage transmission lines. It used expensive generating units located near the load centers to meet the periodic demand peaks throughout the year. This combination of local generation and transmission capacity to deliver distant generation was the least-cost systemwide strategy for serving the utility's total demand in the former regime.

The transmission network that resulted from this strategy by the vertically integrated utility for serving its retail customers creates local market power problems in the new wholesale market regime because now the owner of the generating units located close to the load center may not own, and certainly does not operate, the transmission network. The owner of the local generation units is often unaffiliated with the retailers serving customers in that geographic area. Consequently, during the hours of the year when system conditions require that some energy be supplied from these local generation units, it is profit-maximizing for the owners to bid whatever the market will bear for any energy they provide.

This point deserves emphasis: the bids of the units within the local area must be taken before lower-priced bids from other firms outside this area because the configuration of the transmission network and location of demand makes these units the only ones physically capable of meeting the energy need. Without some form of regulatory intervention, these suppliers must be paid at their bid price in order to be willing to provide the needed electricity. The configuration of the existing

[6]A single price market sets a one price of electricity for the entire market. A zonal-pricing market sets different prices for different geographic regions or zones when there transmission congestion between adjacent zones. A nodal-pricing model sets a different price for each node (withdrawal or injection points in the transmission network) if there are transmission constraints between these nodes.

transmission network and the geographic distribution of generation capacity ownership in all US wholesale markets and a number of wholesale markets around the world results in a frequency and magnitude of substantial local market power for certain market participants that if left unmitigated could earn the generation unit owners enormous profits and therefore cause substantial harm to consumers. Designing regulatory interventions to limit the exercise of local market power is a major market design challenge.

3.2. Regulatory Challenges in Wholesale Market Regime

The primary regulatory challenge of the wholesale electricity market regime is limiting the exercise of unilateral market power by market participants. The explicit exercise of unilateral market power is not possible in the vertically-integrated utility regime because the regulator, not a market mechanism sets the price the firm is allowed charge. This is the primary reason why a wholesale electricity market requires substantially more sophistication and economic expertise from the regulatory process at both the federal and state levels than is necessary under the vertically-integrated utility regime.

The regulatory process in the vertically-integrated utility regime is an administrative procedure that focuses on determining the utility's prudently incurred costs and setting prices for the utility's outputs that recovers only these costs. Until very recently, the use of the regulatory price-setting process to provide incentives for least-cost production and higher service quality as discussed in Paul Joskow's chapter was not considered.

The regulatory process for the wholesale market regime must limit the exercise of unilateral market power in the industry segments where market mechanisms are used to set prices. The regulatory process must also determine the allowed revenues and prudency of investment decisions by the transmission and distribution network owners, the two monopoly segments of the industry. However, different from the vertically-integrated utility regime, these investment decisions can impact wholesale electricity market outcomes. Specifically, the capacity of transmission link can impact the number of independent suppliers able to compete to provide electricity at a given location the transmission network, which exerts a direct influence on wholesale electricity prices.

The major regulatory challenge in the wholesale market regime is how to design market-based mechanisms for the wholesale and retail segments of the industry that cause suppliers to produce in a least-cost manner and set prices that come as close as possible to recovering only

production costs. This is essentially the same goal as the vertically-integrated utility regulatory process, but it requires far more sophistication and knowledge of economics and engineering to accomplish because firms have far greater discretion to foil the regulator's goals through their unilateral actions. They can withhold output from their generation units and offer these generation units into the market at prices that far exceed each unit's variable cost of production in order to the raise market-clearing price. Firms can also use their ownership of transmission assets and financial transmission rights to increase their revenues from participating in the wholesale market. It is very difficult for the regulator to prevent these actions if they are in the unilateral interest of the market participant. The combined federal and state regulatory process must therefore determine what wholesale and retail market rules will make it in the unilateral interest of all market participants to set wholesale and retail prices that allow suppliers and retailers the opportunity to recover their prudently incurred costs. This is essence of the market design problem.

4. Market Design Process

This section provide a theoretical framework for describing the important features of the market design process. It is first described in general terms using a generic principal-agent model. The basic insight of this perspective is that once a market rule is set, market participants maximize their objective functions, typically expected profits for privately-owned market participants, subject to the constraints imposed on their behavior by this market rule. The market designer must therefore anticipate how market participants will respond to any market rule in order to craft a design that ultimately achieves the its objectives. The technology of supplying electricity described in Section 2.2 and the regulatory structure governing the industry described in Section 2.3 also place constraints on the market design process. This section introduces the concept of a residual demand curve to summarize the constraints imposed on each market participant by the market rules, technology of producing electricity, and regulatory structure of the industry and uses it to illustrate the important dimensions of the market design process for wholesale electricity.

For the purposes of this discussion, I assume that the goal of the market design process is to achieve the lowest possible annual average retail price of electricity consistent with the long-term financial viability of the industry. Long-term financial viability of the industry, implies that these retail prices are sufficient to fund the necessary new investment to meet demand growth and replace depreciated assets into the indefinite future. Other goals for the market design process are possible,

but this one seems most consistent with the goal of state-level regulatory oversight in the vertically integrated regime.

4.1. Dimensions of Market Design Problem

There are two primary dimensions of the market design problem. The first is the extent to which market mechanisms versus regulatory processes are used to set the prices consumers pay. The second is the extent to which market participants are government versus privately owned. Given the technologies for producing and delivering electricity to final consumers, the market designer faces two basic challenges. First is how to cause producers to supply electricity in both a technically and allocatively efficient manner. Technically efficient production obtains the maximum amount of electricity for given quantity of inputs, such as capital, labor, materials and input energy. Allocatively efficient production uses the minimum cost mix of inputs to produce a given level of output.

The second challenge is how to set the prices for the various stages of the production process that provide strong incentives for technically and allocatively efficient production yet only recover production costs including a return on the capital invested. This process involves choosing a point in the market versus regulation dimension and government versus private ownership dimension for each segment of the electricity supply industry.

Conceptually, the market designer chooses the number and sizes of each market participant and the rules for determining the revenues received by each market participant to maximize its objective function. There are two key constraints on the market designer's optimization problem implied by the behavior of market participants. The first is that once the market designer chooses the rules for translating a market participant's actions into the revenues it receives, each market participant will choose a strategy that maximizes his payoff given the rules set by the market designer. This constraint implies that the market designer must recognize that all market participants will maximize their profits given the rules the market designer selects. The second constraint is that each market participant must expect to receive from the compensation scheme chosen by the market designer more than their opportunity cost of participating in the market. The first constraint is called the individual rationality constraint because it assumes each market participant will behave in a rational (expected payoff-maximizing) manner. The second constraint

is called the participation constraint, because it implies that firms must find participation in the market more attractive than their next best alternative.

4.2. Generic Principal-Agent Problem

To make these features of the market design problem more concrete, it is useful to consider a simple special case of this process–the generic principal-agent model. Here a single principal designs a compensation scheme for a single agent that maximizes the principal's expected payoff subject to the agent's individual rationality constraint and participation constraint. Let W(x,s) denote the payoff of the principal given the observable outcome of the interaction, x, and state of the world, s. The observable outcome, x, depends on the agent's action, a, and the true state of the world, s. Writing x as the function x(a,s) denotes the fact that it depends on the both of these variables.

Let V(a,y,s) equal the payoff of the agent given the action taken by the agent, a, the compensation scheme set by the principal, y(x), and the state of the world, s. The principal's action is to design the compensation scheme, y(x), a function that relates the outcome observed by the principal, x, to the payment made to the agent.

With this notation, it is possible to define the two constraints facing the principal in designing y(x). The individual rationality constraint on the agent's behavior is that it will choose its action, a, to maximize its payoff V(a,y,s) (or the expected value of this payoff) given y(x) and s (or the distribution of s). The participation constraint implies that the compensation scheme y(x) set by the principal must allow the agent to achieve at least its reservation level of utility or expected utility, V^*.

There are two versions of this basic model. The first assumes that the agent does not observe the true state of the world when it takes its action, and the other assumes the agent observes s before taking its action. In the first case, the agent's choice is:

$$a^* = \underset{a}{argmax}\ E_s(V(a,y(x),s),$$

where $E_s(.)$ denotes the expectation with respect to the distribution of s. The participation constraint is $E_s(V(a^*, y(x^*),s)) > V^*$, where $x^* = x(a^*,s)$, which implies that the agent expects to receive utility greater than its reservation utility. In the second case, the agent's problem is:

$$a^*(s) = \underset{a}{argmax}\ V(a,y(x),s),$$

and the participation constraint is $V(a^*(s), y(x^*), s) > V^*$ for all s, where $x^* = x(a^*(s),s)$ in this case.

An enormous number of bilateral economic interactions fit this generic principal-agent framework. Examples include the client-lawyer, patient-doctor, lender-borrower, employer-worker, and firm owner-manager interactions. A client seeking legal services designs a compensation scheme for her lawyer that depends on the observable outcomes (such as the verdict in the case) that causes the lawyer to maximize the client's expected payoff function subject to constraint the lawyer will take actions to maximize his expected payoff given this compensation scheme and the fact that the lawyer must find the compensation scheme sufficiently attractive to take on the case. Another example is the firm owner designing a compensation scheme that causes the manager to maximize the expected value of the owner's assets subject to the constraint that the firm manager will take actions to maximize her expected payoff given the scheme is in place and the fact that it must provide a higher expected payoff to the manager than she could receive elsewhere.

4.3. Applying the Principal-Agent Model to the Market Design Process

The regulator-utility interaction is a principal-agent model directly relevant to electricity industry re-structuring. In this case, the regulator designs a scheme for compensating the vertically-integrated utility for the actions that it takes recognizing that once this regulatory mechanism is in place the utility will attempt to maximize its payoff function subject to this regulatory mechanism. In this case, y(x), would be the mechanism used by the regulator to compensate the firm for its actions. For example, under a simple *ex post* cost-of-service regulatory mechanism, x would be the output produced by the firm, and y(x) would be the firm's total cost of providing this output. Under a price cap regulatory mechanism, x would be the change in the consumer price index for the US economy and y(x) would be the total revenues the firm receives, assuming it serves all demand at the price set by this regulatory mechanism. The incentives for firm behavior created by any potential regulatory mechanism can be studied within the context of this principal-agent model.

This modeling framework is also useful for understanding the incentives for firm behavior in a market environment. A competitive market is another possible way to compensate a firm for the actions that it takes. For example, the regulator could require this firm and other firms to bid

their willingness to supply as a function of price and only chose the firms with bids below the lowest price necessary to meet the aggregate demand for the product. In this case x can be thought of as the firm's output and y(x) the firm's total revenues from producing x and being paid this market-clearing price per unit sold. Viewed from this perspective, markets are simply another regulatory mechanism for compensating a firm for the actions that it takes.

It is well-known that profit-maximizing firms that are not constrained by a regulatory price-setting process have strong incentive to produce their output in an technically and allocatively efficient manner. However, it is also well-known that profit-maximizing firms have no unilateral incentive to pass on these minimum production costs in the price they charge to consumers. It is only when competition among firms is sufficiently fierce that this will occur.

Economic theory provides conditions under which a market will yield an optimal solution to the problem of causing the suppliers to provide their output to consumers at the lowest possible price. One of these conditions is the requirement that suppliers are atomistic, meaning that all producers believe they are so small relative to the market that they have no ability to influence the market price through their unilateral actions. Unfortunately, this condition is unlikely to hold for the case of electricity given the size of most market participants before the reform process starts. These firms recognize that if they remain large, they will have the ability to influence both market and political outcomes through their unilateral actions. Moreover, the minimum efficient scale of electricity generation, transmission and distribution is such that it is unlikely to be least cost for the industry as a whole to separate electricity production into a large number of extremely small firms. So there is an underlying economic justification for allowing these firms to remain large, although certainly not as large as they would like to be. This is one reason why the electricity market design process is so difficult. This problem is particularly acute for small countries or regions without substantial transmission interconnections with neighboring countries or regions.

This principal-agent model is also useful for understanding why industry outcomes can differ so dramatically depending on whether the industry is government or privately owned. First, the objective function of the firm's owner differs across the two regimes. Under government ownership all of the citizens of the country are shareholders. These owners are also severely limited in the sorts of mechanisms they can design to compensate the management of the firm. For example, there is no liquid market for selling their ownership stake in this firm. It is virtually impossible for them to

remove the management of this firm. They don't even have a legal right to their ownership stake in the firm. In contrast, a shareholder in a privately-owned firm has a clearly defined and legally enforceable property right than can be sold in a liquid market. If they own enough shares in the firm or can get together with other large shareholders, they can remove the management of the company. Finally, by selling their shares, they can severely limit the ability of the company to raise capital for new investment. In contrast, the government-owned firm obtains the funds necessary for new investment primarily through the political process.

This discussion illustrates the point that despite the fact that both the government-owned and privately-owned firm have access to exactly the same technologies to generate, transmit and distribute electricity, dramatically different industry outcomes in terms of the mix of generation capacity installed, the price consumers pay and the amount they consume can occur because the schemes for compensating each firm's management, $y(x)$, differ because the owners of the two firms have different objective functions and different sets of feasible mechanisms for compensating their management. Applying the generic principal-agent model to the issue of government versus private ownership implies that different industry outcomes should occur if a government-owned vertically-integrated geographic monopolist is asked to provide electricity to the same geographic area that a privately-owned geographic monopolist previously served, even if both monopolists face the same regulatory mechanism for setting the prices they charge to retail consumers.

Applying the logic of the principal-agent model at the level of the regulator-firm interaction as opposed to the firm owner-management interaction implies an additional source of differences in market outcomes if, as is often the case, the government-owned monopoly faces a different regulatory process than the privately-owned monopoly. Laffont and Tirole (1991) build on this basic insight to construct a theoretical framework to study the relative advantages of public versus private ownership. They formulate a principal-agent model between the management of the publicly-owned firm and the government in which the cost of public ownership is "suboptimal investment by the firm's managers in those assets that can be redeployed to serve the goals pursued by the public owners" (Laffont and Tirole, 1991, p. 84). The cost of private ownership in their model is the classical conflict between the desire of the firm's shareholders for it to maximize profits and the regulator's desire to limit these profits. Laffont and Tirole (1991) find that the existence of these two agency relationships does not allow a general prediction about the relative social

efficiency of public versus private ownership, although the authors are able characterize circumstances where one ownership form would dominate the other.

In the wholesale market regime, the extent of government participation in the industry creates an additional source of differences in industry outcomes. As Laffont and Tirole (1991) argue, the nature of the principal-agent relationship between the firm's owner and its management is different under private ownership versus government ownership. Consequently, an otherwise identical government-owned firm can be expected to behave differently in a market environment from how this firm would behave if it were privately owned. This difference in firm behavior yields different market outcomes depending on the ownership status (government versus privately-owned) of the firms in the market.

Consequently, in its most general form, the market design problem is composed of multiple layers of principal-agent interactions where the same principal can often interact with a number of agents. For example, the case of a competitive wholesale electricity market, the same regulator interacts with all of firms in the industry. The market designer must recognize the impact of all of these principal-agent relationships in designing an electricity supply industry to achieve his market design goals. The vast majority of electricity market design failures result from ignoring the individual rationality constraints implied by both the regulator-firm and firm owner-management principal-agent relations. The individual rationality constraint most often ignored is that privately-owned firms will maximize their profits from participating in a wholesale electricity market. It is important to emphasize that this individual rationality constraint holds whether or not the privately-owned profit-maximizing firm is one of a number of firms in a market environment or a single vertically integrated monopolist. The only difference between these two environments is the set of actions that the firm is legally able to take to maximize its profits.

4.4. Individual Rationality Under a Market Mechanism versus a Regulatory Process

The set of actions available to firms in a market environment is different from those available to it in a regulated-monopoly environment. For example, under a market mechanism firms can increase their profits by both reducing the costs of producing a given level of output or by increasing the price they charge for this output. In contrast, under the regulated-monopoly environment, the firm does not set the price it receives for its output. Instead, the legal contract between the firm and regulator requires the firm to supply all that is demanded at a price set by the regulator in exchange

for the firm being given a legal monopoly to supply a given geographic area and the opportunity to earn a reasonable rate return on their investment from the prudent operation of their facilities and selling their output at the price set by the regulator.

Defining the incentive constraint for a privately-owned firm operating in a competitive electricity market is relatively straightforward. Because the firm would like to maximize profits, it has a strong incentive to produce its output at minimum cost. In other words, the firm will produce in a technically and allocatively efficient manner. However, the firm has little incentive to set a price that only recovers these production costs. In fact, the firm would like to take actions to raise the price it receives above both the average and marginal cost of producing its output. Profit-maximizing behavior implies that the firm will choose a price or level of output such that the increase in revenue it earns from supplying one more unit equals the additional cost that it incurs from producing one more unit of output. This is the same thing as saying that the firm will withhold output from the market until the cost savings from withholding one more unit of output is less than or equal to the total revenue loss from withholding that unit of output from the market.

Figure 8 provides a simple model of the unilateral profit-maximizing behavior of a supplier in a bid-based electricity market. Let Q_d equal the level of market demand for a given hour and SO(p) the aggregate willingness to supply as a function of price of all other market participants besides the firm under consideration. Figure 8(a) plots the inelastic aggregate demand curve and the upward sloping supply of all other firms besides the one under consideration. Figure 8(b) subtracts this aggregate supply curve of all other market participants from the market demand to produce to the residual demand curve faced by this supplier, $DR(p) = Q_d - SO(p)$. This panel also plots the marginal cost curve for this supplier, as well as a the marginal revenue curve associated with DR(p).

The intersection of this marginal revenue curve with the supplier's marginal cost curve yields the profit-maximizing level of output and market price for this supplier given the bids submitted by all other market participants. This price-quantity pair is denoted by (P^*,Q^*) in Figure 8(b). Profit-maximizing behavior by the firm implies the following relationship between the marginal cost at Q^*, which I denote by $MC(Q^*)$, and P^* and ε, the elasticity of the residual demand at P^*:

$$(P^* - MC(Q^*))/P^* = - 1/\varepsilon, \qquad (1)$$

where $\varepsilon = DR'(P^*)*(P^*/DR(P^*))$. Because the slope of the firm's residual demand, $DR'(P^*)$, at this level of output is finite, the market price is larger than supplier's marginal cost. The price-quantity pair associated with the intersection of DR(p) with the supplier's marginal cost curve is denoted (P^c,Q^c). It is important to emphasize that even though the price-quantity pair (P^c,Q^c) is often called the competitive output level, producing at this level is not unilateral profit-maximizing for the firm if it faces a downward sloping residual demand curve. This is another way of saying that price-taking behavior–acting as if the firm had no ability to impact the market price–is never individually rational. It will only occur as an equilibrium outcome if competitive conditions in the market are particularly fierce.

A firm that influences market prices as shown in Figure 8(a)-(b) is said to be exercising its unilateral market power. A firm possesses unilateral market power if has the ability to raise the market price through its unilateral actions and profit from this price increase. We would expect all privately-owned profit-maximizing firms to behave in this manner. This is equivalent to saying that the firm satisfies its individual rationality constraint. I would like to emphasize that as long as a supplier faces a residual demand curve with any upward slope, it has the ability to exercise unilateral market power.

In virtually all oligopoly industries, the best information a researcher can hope to observe is the market-clearing price and quantity sold by each firm. However, in a bid-based wholesale electricity market much more information is typically available to the analyst. The entire residual demand curve faced by a supplier, not just a single point, can be computed the using bids and offers of all other market participants. The market demand Q_d is observable and the aggregate willingness to supply curve of all other firms besides the one under consideration, SO(p), can be computed from the willingness-to-supply offers of all firms. Therefore, it is possible to compute the elasticity of residual demand curve for any price level including the market-clearing price P^*. The absolute value of the inverse of the elasticity of the residual demand curve, $|1/\varepsilon|$, for $\varepsilon = DR'(P^*)*(P^*/DR(P^*))$, measures the percentage increase in the market-clearing price that would result from the firm under consideration reducing its output by one percent. Note that this measure depends on the level of market demand and the aggregate willingness-to-supply curve of the firm's competitors. Therefore, this inverse elasticity of the residual demand curve measures the firm's ability to raise market prices

through its unilateral actions (given the level of market demand and the willingness to supply offers of its competitors).

Figure 8(c)-(d) illustrates the extremely unlikely case that the supplier faces an infinitely elastic residual demand curve and therefore finds it in its unilateral profit-maximizing to produce at the point that the market price is equal to its marginal cost. This point is denoted (P^{**},Q^{**}). The supplier faces an infinitely elastic residual demand curve because the SO(p) curve is infinity elastic at P^{**}, meaning that all other firms besides this supplier are able produce all that is demanded if the price is above P^{**}. Note that even in this extreme case the supplier is still satisfying the individual rationality constraint by producing at the point that the marginal revenue curve associated with DR(p), crosses its marginal cost curve, as is required by equation (1). The only difference is that this marginal revenue curve is associated with this residual demand curve also equal to the supplier's average revenue curve, because DR(p) is infinitely price elastic, meaning that it is a horizontal line. Because the slope of the firm's residual demand curve is infinite, $1/\varepsilon$, is equal to zero which implies that the firm has no ability to influence the market price through its unilateral actions and will therefore find unilaterally profit-maximizing to produce at the point that the market-clearing price equals its marginal cost.

Figure 8 demonstrates that the individual rationality constraint in the context of a market mechanism is equivalent to the supplier exercising all available unilateral market power. Even in the extreme case of the infinitely elastic residual demand curve in Figure 8(d), the supplier still exercises all available unilateral market power. However, in this case the supplier cannot increase its profits by withholding output that can be produced at a marginal cost less than market price, because it has no ability to exercise unilateral market power.

Individual rationality in the context of a regulatory process still implies that the firm will maximize profits given the mechanism for compensating it set by the regulator. However, in this case the firm is unable to set the price it charges consumers or the level of output it is willing to supply. The firm must therefore take a more subtle approaches to maximizing its profits because the regulator sets the output price and requires the firm to supply all that is demanded at this regulated price. In this case the individual rationality constraint can imply that the firm will produce its output in a technically or allocatively inefficient manner because of how the regulatory process sets the price that the firm is able to charge.

The well-known Averch and Johnson (1962) model of cost-of-service regulation assumes that the regulated firm produces its output using capital, K, and labor, L, yet the price the regulator allows the firm to charge for capital services is greater than the actual price the regulated firm pays for capital services. This implies that a profit-maximizing firm facing the zero-profit constraint implied by this regulatory process will produce its output using capital more intensively relative to labor than would be the case if the regulatory process did not set a price for capital services different from the one the firm actually pays. The Averch and Johnson model illustrates a very general point associated with the individual rationality constraint in regulated settings: It is virtually impossible to design a regulatory mechanism that causes a privately-owned profit-maximizing firm to produce in an least-cost manner if the firm's output price is set by the regulator based on its incurred production costs.

The usual reason offered for why the regulator is unable to set prices that achieve the market designer's goal of least cost production is that the regulated firm usually knows more about its production process or demand than the regulator. Although both the firm and regulator have substantial expertise in the technology of generating, transmitting and distributing electricity to final consumers, the firm has a much better idea of precisely how these technologies are implemented. This informational asymmetry leads to disputes between the firm and the regulator over the minimum cost mode of production to serve the firm's demand. Consequently, the regulator can never know the minimum cost mode production to serve final demand.

Moreover, there are laws against the regulator confiscating the firm's assets through the prices it sets, and the firm is aware of this fact. This creates the potential for disputes between the firm and the regulator over the price level that provides strong incentives for least-cost production, but does not confiscate the firm's assets. All governments recognize this fact and allow the firm an opportunity to subject a decision by the regulator about the level of the firm's output price to judicial review. To avoid the expense and potential loss of credibility of a judicial review, the regulator may instead prefer to set a slightly higher regulated price to guarantee that the firm will not appeal its decision. This aspect of the regulatory process reduces the incentive the firm has to produce its output in a least cost manner.

Wolak (1994) is an empirical study of the regulator-utility interaction between California water utilities and the CPUC. He specifies and estimates an econometric model of this principal-

agent interaction and quantifies the magnitude of the distortions from minimum cost production induced by the informational asymmetries between firm and the regulator about the utility's production process. Even for the relatively simple technology of providing local water delivery services, where the extent of informational asymmetries between the firm and the regulator are likely to be small, Wolak (1994) finds that actual production costs are between 5% and 10% higher than they would be under least cost production. Deviations from least-cost production in a vertically-integrated electricity supply industry are likely to be much greater because the extent of the informational asymmetries between the firm and regulator about the firm's production process are likely to be much greater than in the water distribution industry. The substantially greater complexity of the process of generating and delivering electricity to final consumers implies more sources of informational asymmetries between the firm and regulator and therefore the potential for greater distortions from least-cost production.

The market designer does not need to worry about the impact of informational asymmetries on a firm's mode of production in a competitive market. There is no legal requirement that the market set the price the firm is paid for its output above some minimum level. Different from regulated environments, there are no laws against a competitive market setting prices that confiscate a firm's assets. Any firm that is unable to cover its costs of production at the market price must eventually exit the industry. Firms cannot file for a judicial review of the prices set by a competitive market. Competition among firms leads high-cost firms to exit the industry and be replaced by lower cost firms. Contrary to the regulated regime, there is no need to determine if a firm's incurred production costs are the result of the least-cost mode of production. If the market is sufficiently competitive and has low barriers to entry, then any firm that is able to remain in business must be producing its output at or close to minimum cost. Otherwise a more efficient firm could enter and profitably underprice this firm. The risk that firms not producing in a least cost manner will be forced to exit creates much stronger incentives for least-cost production than would be the case under regulation, where the firm recognizes that the regulator does not know the least-cost mode of production and can exploit this fact through less technically and allocatively inefficient production that may ultimately yield the firm higher profits.

This difference in the incentives for least-cost production under regulation versus a market mechanism reinforces the impact of individual rationality constraints on firm behavior under a

market regime versus a regulated utility regime. In the case of a market mechanism, the individual rationality constraint provides strong incentives for each firm to produce its output at least cost, but little, if any, incentive to price this output to only recover production costs. In fact, depending on the extent of competition the firm faces, it may have an extremely strong incentive to price its output vastly in excess of the marginal cost of producing the most expensive unit sold. For the case of the price-regulated vertically-integrated utility regime, the individual rationality constraint implies that firm does not produce its output in a least cost manner. Because the regulator sets the price the firm is able to charge and this price is set to only recover the firm's prudently incurred costs, the firm has an incentive to deviate from the least cost mode of production to exploit its superior information.

Consequently, the advantage of regulation is that the market price should not deviate significantly from actual average cost of producing the firm's output. However, the firm has very little incentive to make its actual mode of production equal to the least-cost mode of production. In contrast, the competitive regime provides very strong incentives for firms produce in a least-cost manner. Unless the firm faces sufficient competition, it has little incentive to pass on only these efficiently incurred production costs in the prices charged to consumers. This discussion shows that the potential exists for consumers to pay lower prices under either regime. Regulation may be favored if the market designer is able to implement a regulatory process that is particularly effective at causing the firm to produce in a least-cost manner, or if the market designer is unable to establish a sufficiently competitive market so that prices are vastly in excess of the marginal cost of producing the last unit sold. Competition is favored if regulation is particularly ineffective at providing incentives for least-cost production or competition is particularly fierce. Nevertheless, in making the choice between a market mechanism and a regulatory mechanism, the market designer must typically make a choice between two imperfect worlds–an imperfect regulatory process or an imperfectly competitive market. Which mechanism should be selected depends on which one maximizes the market designer's objective function.

4.5. Individual Rationality Constraint Under Government versus Private Ownership

The individual rationality constraint for a government-owned firm is difficult to characterize for two reasons. First, it is unclear what control the firm's owners are able to exercise over the firm's management and employees. Second, it is also unclear what the objective function of the firm's owners is. For the case of privately-owned firms, there are well-defined answers to both of

these questions. The firm's owners have clearly-specified legal rights and their ownership shares can be bought and sold by incurring modest transactions costs. Because, keeping all other things equal, investors would like to earn the highest possible return on their investments, shareholders would like the firm's management to maximize the risk-adjusted rate of return on equity. This implies that the firm's owners will attempt to devise a compensation scheme for the firm's management that causes them to maximize profits. In comparison, it is unclear if the government wants its firms to maximize profits. Earning more revenues than costs is clearly a priority, but once this is accomplished the government would most likely want to the firm to pursue other goals. This is the tension that Laffont and Tirole (1991) introduce into their model of the behavior of publicly-owned firms.

This lack of clarity in the both the objective function of the government for the firms it owns and the set of feasible mechanisms the government can implement to compensate the firm's management has a number of implications. The first is that it is unlikely that the management of a government-owned firm will produce and sell its output in a profit-maximizing manner. Different from a privately-owned firm, its owners are not demanding the highest possible return on their equity investments in the firm. Because a government-owned firm's management has little incentive to maximize profits, it also has little incentive to produce in a least-cost manner. This logic also implies that a government-owned firm has little incentive to attempt to raise prices beyond the level necessary to cover its total costs of production. The second implication of this lack of clarity in objectives and feasible mechanisms is that the firm's management now has the flexibility to pursue a number of other goals besides minimizing the total cost of producing the output demanded by consumers.

Viewed from the perspective of the overall market design problem, one advantage of government-ownership is that the pricing goals of the firm do not directly contradict the market designer's goal of the lowest possible prices consistent with the long-term financial viability of the industry. In the case of private-ownership, the pricing incentives of the firm's management directly contradict the interests of consumers. The firm's management wants to raise prices above the marginal cost of the last unit produced, because of the desire of the firm's owner to receive the highest possible return on their investment in the company. The desire of privately-owned firms to maximize profits leads to pricing incentives that directly contradict the goals of the market design

process. Unless the firm faces a sufficient competition from other suppliers, which from the discussion of Figure 8, is equivalent to saying that the firm faces a sufficiently elastic residual demand curve, this desire to maximize profits will yield market outcomes that reflect the exercise of significant unilateral market power.

However, it is important to emphasize that prices set by a government-owned firm may cause at least as much harm to consumers as prices that reflect the exercise of unilateral market power if the incentives for least-cost production by the government-owned firm are sufficiently muted and the firm is required to set a price that at least recovers all of its incurred production costs. Although these prices may appear more benign because they only recover the actual costs incurred by the government-owned firm, they can be more harmful from a societal welfare perspective than the same level of prices set by a privately-owned firm. This is because the privately-owned firm has a strong incentive to produce in a technically and allocatively efficient manner and any positive difference between total revenues paid by consumers and the minimum cost of producing the output sold is economic profit or producer surplus.

Government-owned firms may produce in technically and/or allocatively inefficient manner because of constraints imposed by its owner. For example, the government could require a publicly-owned firm to hire more labor than is necessary. This is socially wasteful and therefore yields a reduced level of producer surplus relative to case of a privately-owned firm producing its output in a least-cost manner. Because both outcomes, by assumption, have consumers paying the same price, the level of consumer surplus is unchanged across the two ownership structures, so that the level of total surplus is reduced as a result of government-ownership because the difference between the market price and the variable cost of the highest cost unit operating under private ownership goes to the firm's shareholders in the form of higher profits.

Figure 9 provides a graphical illustration of this point. The step function labeled MC_p is the incurred marginal cost curve for the privately-owned firm and step function labeled MC_g is incurred marginal cost curve for the government-owned firm. I make the distinction between incurred and minimum cost to account for the fact that the management of the government owned-firm has less of an incentive to produce at minimum cost than does the privately-owned firm. In this example, I assume the reason for this difference in marginal cost curves is that the government-owned firm produces in a technically inefficient manner by using more of each input to produce the same level

of output as the privately-owned firm. Suppose that the profit-maximizing level of output for the privately-owned firm given the residual demand curve plotted in Figure 9 is Q^*, with a price of P^*. Suppose the government-owned firm behaves as if it were a price-taker given its marginal cost curve and this residual demand curve and assume that this price is also equal to the firm's average incurred cost at Q^*, $AC(Q^*)$. I have drawn the figure so that the intersection of the marginal cost curve of the government-owned firm with this residual demand curve occurs at the same price and quantity pair set by the unilateral profit-maximizing quantity offered by the privately-owned firm.

Because the government-owned firm produces in a technically inefficient manner, it uses more of society's scarce resources to produce Q^* than the privately-owned firm. Consequently, the additional benefit that society receives from having the privately-owned firm produce the good is the shaded area between the two marginal cost curves in Figure 9, which is the additional producer surplus earned by the privately-owned firm because it produces in a technically and allocatively efficient manner but exercises significant unilateral market power.

This example demonstrates that even though the privately-owned firm exercises all available unilateral market power, if the incentives for efficient production by government-owned firms are sufficiently muted, it may be preferable from the market designer's and society's perspective to tolerate some exercise of unilateral market power, rather than adopt a regime with government-owned firms setting prices equal to an extremely inefficiently incurred marginal cost or average cost of production.

If the government-owned firm is assumed to produce in an allocatively inefficient manner only, this same logic for consumers preferring private to government ownership holds. However, the societal welfare implications of government-versus private ownership are less clear because these higher production costs are caused by deviations from least-cost production rather than simply a failure to produce the maximum technically feasible output for a fix set of inputs. For example, if the government-owned firm is forced to pay higher wages than private sector firms for equivalent workers because of political constraints, these workers from the government-owned firm would suffer a welfare loss if they were employed by a privately-owned firm.

The example given in Figure 9 may seem extreme, but there are number of reasons why it is reasonable to believe that a government-owned firm faces far less pressure from its owners to produce in a least cost manner relative to its privately-owned counterpart. For example, poorly run

privately-owned companies can go bankrupt. If a firm consistently earns revenues less than its production costs, the firm's owners and creditors can force the firm to liquidate its assets and exit the industry. The experience from both industrialized and developing countries is that poorly run government-owned companies rarely go out of business. Governments can and almost always do fund unprofitable companies from general tax revenues. Even in the US, there are a number of examples of persistently unprofitable government-owned companies receiving subsidies long after it is clear to all independent observers that these firms should liquidate their assets and exit the industry. Because government-owned companies have this additional source of funds to cover their incurred production costs, they have significantly less incentive to produce in a least-cost manner.

Megginson and Netter (2001) survey a number of empirical studies of the impact of privatization in non-transition economies and find general support for the proposition that it improves the firm's operating and financial performance. However, these authors are emphasize that this improved financial performance does not always translate into increases in consumer welfare because private ownership can increase the incentive for firms to exercise unilateral market power. Shirley and Walsh (2000) also survey the empirical literature on the impact of privatization on firm performance. They conclude that the private ownership and competition are complements in the sense that the empirical evidence on private ownership improving firm performance is stronger when the private firm faces competition. They also argue that the relative performance improvements associated with private versus public ownership are greater in developing countries versus industrialized countries.

5. Dimensions of Wholesale Market Design Process

This section describes the five major ways that a market designer can reduce the incentive a supplier has to exercise unilateral market power in a wholesale electricity market. As discussed previously, it is impossible to eliminate the ability that suppliers in a wholesale electricity market have to exercise unilateral market power. The best that a market designer can hope to do is reduce this incentive to levels that yield market outcomes that come closer to achieving the market designer's goals than could be achieved with other feasible combinations of market and regulatory mechanisms. This means the market designer must recognize the individual rationality constraint that the firm will maximize profits given the market rules set by the market designer and the actions taken by the its competitors.

As the discussion of Figure 8 demonstrates, the market designer reduces the ability of the firm to exercise unilateral market by facing the firm with a residual demand curve that is as elastic as possible. As Figure 8 demonstrates, the more elastic the supplier's residual curve demand is the less the firm's unilateral profit-maximizing actions are able to raise the market-clearing price. Consequently, the goal of designing a competitive electricity market is straightforward: Face all suppliers with as elastic as possible residual demand curves during as many hours of the year as possible.

There are five primary mechanisms for increasing the elasticity of the residual demand curve faced by a supplier in a wholesale electricity market. The first is divestiture of capacity owned by this firm into a larger number of independent suppliers. Second is the magnitude and distribution across suppliers of fixed-price forward contracts to supply electricity to sold load-serving entities. Third is the extent to which final consumers are active participants in the wholesale electricity market. Fourth is the extent to which the transmission network has enough capacity to face each supplier with sufficient competition from other suppliers. The last is the extent to which regulatory oversight of the wholesale market provides strong incentives for all market participants to fulfill their contractual obligations and obey the market rules. We now discuss each of these mechanisms for increasing the elasticity of the residual demand curve facing a supplier.

5.1. Divestiture of Suppliers

To understand how the divestiture of a given amount of capacity into a larger number of independent suppliers can impact the slope of a firm's residual demand curve, consider the following simple example. Suppose there are ten equal sized firms, each of which owns 1,000 MW of capacity and that the total demand in the hourly wholesale market is equal to 9,500 MWh. Each firm knows that at least 500 MW of its capacity is needed to meet this demand, regardless of the actions of its competitors. Specifically, if the remaining 9 firms bid all 1,000 MW of the their capacity into the market, the tenth firm has a residual demand of at least 500 MWh at every bid price. Mathematically, this means the value of the residual demand facing the firm, DR(p), is positive at p_{max}, the highest possible bid price that a supplier can submit. When $DR(p_{max}) > 0$, the firm is said to be pivotal, meaning that at least $DR(p_{max})$ of its capacity is needed to serve demand. Figure 10 provides an example of this phenomenon. Let $SO_1(p)$ represent the aggregate willingness-to-supply curve of all other firms besides the firm under consideration and Q_d the market demand. Figure

10(b) shows that the firm is pivotal for $DR_1(p_{max})$ units of output, which in this example is equal to 500 MWh. In this circumstance, the firm is guaranteed total revenues of at least $DR_1(p_{max})*p_{max}$, which it can achieve by bidding all of its capacity into the wholesale market at p_{max}.

To see the impact of a requiring a firm to divest generation capacity on the form of its residual demand curve, suppose that the firm in Figure 10 was forced to sell off 500 MW of its capacity to a new or existing market participant. This implies that the maximum supply of all other firms is now equal to 9,500 MWh, the original 9,000 MWh plus the additional 500 MWh divested, which is exactly equal to the level of demand. This means that the firm is no longer pivotal because, its residual demand is equal to zero at p_{max}. Figure 10(a) draws new bid supply curve of all other market participants besides the firm under consideration, $SO_2(p)$. For every price, I would expect this curve to lie to the right of $SO_1(p)$, the original bid supply curve. Figure 10(b) plots the resulting residual demand curve for the firm using $SO_2(p)$. This residual demand curve, $DR_2(p)$, crosses the vertical axis at p_{max}, so that the elasticity of the residual demand curve facing the firm is now finite for all feasible prices. In contrast, for the case of $DR_1(p)$, the residual demand pre-divestiture, the firm faces a demand of at least $DR_1(p_{max})$ for all prices in the neighborhood of p_{max}.

This example illustrates a general phenomenon associated with structural divestiture, the firm that sells generation capacity now faces a more elastic residual demand curve, which causes it to bid more aggressively into the wholesale electricity market. This more aggressive bidding by the divested firm then faces all other suppliers with flatter residual demand curves, so they now find it optimal to submit flatter bid supply curves, which implies a flatter residual demand curve for the firm under consideration. Even in those cases when divestiture does not stop a supplier from being pivotal, the residual demand curve facing the firm that now has less capacity should still be a more elastic, because more supply has been added to SO(p), the aggregate bid supply function of all other firms besides the firm under consideration. This implies a smaller value for the firm's residual demand at all prices, as shown in Figure 10.

This residual demand analysis illustrates why it is preferable to divest capacity to new entrants or small existing firms rather than to large existing firms. Applying the reverse of the logic described above to the existing supplier than purchases the divested capacity, implies that this firm faces a residual demand that is likely to be smaller at every price level. The acquiring firm now owns generation capacity that formerly had a willingness-to-supply curve that entered acquiring

firm's residual demand curve. The larger amount of generation capacity owned by the acquiring firm before the divestiture occurs, the greater are the likely competition concerns associated with this acquisition.

5.2. Fixed-Price Forward Contracts and Vesting Contracts

In many industries wholesalers and retailers sign fixed-price forward contracts to manage the risk of spot price volatility. There are two additional reasons for wholesalers and retailers to sign fixed-price forward contracts in the electricity supply industry. First, fixed-price forward contract commitments make it unilaterally profit-maximizing for a supplier to submit bids into the spot electricity market closer to its marginal cost of production. This point is demonstrated in detail in Wolak (2000a). Second, fixed-price forward contracts can also pre-commit generation unit owners to a lower average cost pattern of output throughout the day. This logic implies that for the same sales price, a supplier with a significant fixed-price forward contract commitments earns a higher per unit profit than one with a lower quantity of fixed-price forward contract commitments. Wolak (2007b) demonstrates the empirical relevance of this point for a large supplier in the Australian electricity market.

To understand the impact of fixed-price forward contract commitments on supplier bidding behavior it is important to understand what a forward contract obligates a supplier to do. Usually fixed-price forward contracts are signed between suppliers and load-serving entities. These contracts typically give the load-serving entity the right to buy a fixed quantity of energy at a given location at a fixed price. Viewed from this perspective, a forward contract for supply of electricity obligates the seller to provide insurance against short-term price volatility at a pre-specified location in the transmission network for a pre-specified quantity of energy. The seller of the forward contract does not have to produce energy from its own generating facilities to provide this price insurance to the purchaser of the forward contract. However, one way for the seller of the fixed-price forward contract to limit its exposure to short-term price risk is to provide the contract quantity of energy from its own generation units. If the short-term price at the location the generation unit owner injects energy is the same as the short-term price at the location that the fixed-price forward contract clears against, the firm will earn the difference between the forward contract price, PC, and its marginal cost, MC, times the contract quantity, QC, in variable profits (revenues in excess of variable costs) from the forward contract.

This logic leads to another extremely important point about forward contracts that is not often fully understood by participants in a wholesale electricity market. Delivering electricity from a seller's own generation units is not always a profit-maximizing strategy given the supplier's forward contract obligations. This is also the reason why forward contracts provide strong incentives for suppliers to bid more aggressively (supply functions closer to the generation unit owner's marginal cost function) into the short-term wholesale electricity market.

To see these points, consider the following example taken from Wolak (2000a). Let DR(p) equal the residual demand curve faced by the supplier with the forward contract obligation QC at a price of PC and a marginal cost of MC. For simplicity, I assume that the firm's marginal cost curve is constant, but this simplification does not impact any of the conclusions from the analysis. The firm's variable profits for this time period are:

$$\pi(p) = (DR(p)-QC)(p - MC) + (PC - MC)QC. \quad (2)$$

The first term in (2) is equal to the profit or loss the firm earns from buying or selling energy in the short-term market at a price of p. The second term in (2) is the variable profits the firm earns from selling QC units of energy in the forward market at price PC. The firm's objective is to bid into the spot market in order to set a market price, p, that maximizes $\pi(p)$. Because forward contracts are, by definition, signed in advance of the operation of the short-term market, from the perspective of bidding into the short-term market, the firm treats (PC - MC)QC as a fixed payment it will receive regardless of the short-term price, p. Consequently, the firm can only impact the first term through its bidding behavior in the short-term market.

A supplier with a forward contract obligation of QC, has a very strong incentive to submit bids that set prices below its marginal cost if it believes that DR(p) will be less than QC. This is because the supplier is effectively a net buyer of QC - DR(p) units of electricity, because it has already sold QC units in a forward contract. Consequently, it is profit-maximizing for the firm to want to purchase this net demand at the lowest possible price. It can either do this by producing the power from its own units at a cost of MC or purchasing the additional energy from the spot market. If the firm can push the market price below its marginal cost, then it is profit-maximizing for the firm to meet its forward obligations by purchasing power from the spot market rather paying MC to produce it. Consequently, if suppliers have substantial forward contract obligations, then they have extremely strong incentives to keep market prices very low until the level of energy they

actually produce is greater than their fixed-price forward contract quantity.

The competition-enhancing benefits of forward contract commitments from suppliers can be seen more easily by defining $DR_C(p) = DR(p) - QC$, the net-of-forward-contract residual demand curve facing the firm and $F = (PC - MC)QC$, the variable profits from forward contract sales. In terms of this notation the firm's variable profits becomes $\pi(p) = DR_C(p)(p - MC) + F$, which has exactly the same structure (except for F) as the firm's variable profits from selling electricity if it has no forward contract commitments. The only difference is that $DR_C(p)$ replaces $DR(p)$ in the expression for the supplier's variable profits. Consequently, profit-maximizing behavior implies that the firm will submit bids to set a price in the spot market that satisfies equation (1) with DR(p) replaced by $DR_C(p)$. This implies the following relationship between P^c, the ex post profit-maximizing price, the firm's marginal cost of production, MC, and ε^c, the elasticity of the net-of-forward-contract-quantity residual demand curve evaluated at P^c:

$$(P^c - MC)/P^c = - 1/\varepsilon^c, \tag{3}$$

where $\varepsilon^c = DR_C'(P^c)*(P^c/DR_C(P^c))$.

The inverse of the elasticity of net-of-forward-contract residual demand curve, $1/\varepsilon^c$, is a measure of the incentive (as opposed to ability) a supplier has to exercise unilateral market power. If the firm has some fixed-price forward contract obligations, then a given change in the firm's residual demand as a result of a one percent increase in the market price implies a much larger percentage change in the firm's net-of-forward-contract-obligations residual demand. For example, suppose that a firm is currently selling 100 MWh, but has 95 MWh of forward contract obligations. If a one percent increase in the market price reduces the amount that the firm sells by 0.5 MWh, then the elasticity of the firm's residual demand is -0.5 = (0.5 percent quantity reduction) ÷ (1 percent price increase). The elasticity of the firm's residual demand net of its forward contract obligations is -10 = (10 percent net of forward contract quantity output reduction) ÷ (1 percent price increase). Thus, the presence of fixed-price forward contract obligations implies a dramatically diminished incentive to withhold output to raise short-term wholesale prices, despite the fact that the firm has a significant ability to raise short-term wholesale prices through its unilateral actions. In general, ε^c and ε are related by the following equation:

$$\varepsilon^c = \varepsilon[DR(p)/(DR(p)-QC)].$$

The smaller a firm's exposure to short-term prices–the difference between DR(p) and QC–the more

elastic ε^c is relative to ε, and the greater the divergence between the incentive versus ability the firm has to exercise unilateral market power.

Because $DR_C(p) = DR(p) - QC$, this implies that at same market price, p, and residual demand curve, DR(p), the absolute of value of the elasticity of the net-of-forward-contract-quantity residual demand curve is always greater than the absolute value of the elasticity of the residual demand curve. A simple proof of this result follows from the fact that $DR_C'(p) = DR'(p)$ for all prices and $QC > 0$, so that by re-writing the expressions for ε^c and ε, we obtain:

$$| \varepsilon^c | = | DR'(p)*(p/[DR(p) - QC]) | > | \varepsilon | = | DR'(p)*(p/DR(p)) |. \quad (4)$$

Moreover, as long as $DR(p) - QC > 0$, the larger the value of QC, the greater is the difference between ε^c and ε, and the smaller is the expected profit-maximizing percentage mark-up of the market price above the firm's marginal cost of producing the last unit of electricity that it supplies with forward contract commitments versus no forward contract commitments. This result demonstrates that it is always unilateral profit-maximizing, for the same underlying residual demand curve, for the supplier to set a lower price relative to its marginal cost if it has forward contract commitments.

This incentive to bid more aggressively in the spot market if a supplier has substantial forward contracts also has implications for how a fixed quantity of forward contract commitments should be allocated among suppliers to maximize the benefits of these contracts to the competitiveness of the spot market. Because a firm with forward contract obligations will bid more aggressively in the spot market, this implies that all of its competitors will also face more elastic residual demand curves and therefore find it unilaterally profit-maximizing to bid more aggressively in the spot market. This more aggressive bidding will leave all other firms will more elastic residual demand curves, which should therefore make these firms bid more aggressively in the spot market.

This virtuous cycle with respect to the benefits of forward contracting implies that a given amount of fixed-price forward contracts will have the greatest competitive benefits if it spread out among all of the suppliers in the market roughly proportion to their generation capacity ownership shares. For example, if there are five firms and each them owns 1000 MW of capacity then fixed-price forward contract commitments should be allocated equally across the firms to maximize the competitive benefits. If one firm owned twice the capacity of other firms, then it should have roughly twice the forward contract commitments to load-serving entities that the other suppliers

have.

Because of the spot market efficiency benefits of substantial amounts of fixed-price forward contract commitments between suppliers and load-serving entities, most wholesale electricity market begin operation with a large fraction of the final demand covered by fixed-price forward contracts. If a substantial amount of capacity is initially controlled by government-owned or privately-owned monopolies, the regulator or market designer usually requires that most of these assets be sold to new entrants to create a more competitive wholesale market. These sales typically take place with a fixed-price forward contract commitment on the part of the new owner of the generation capacity to supply a substantial fraction of the expected output of the unit to electricity retailers at a fixed price. These contracts are typically called vesting contracts, because they are assigned to the unit as pre-condition for its sale. For example, if a 500 MW unit owned by the former monopolist is being sold, the regulator assigns a forward contract obligation on the new owner to supply 400 MW of energy each hour at a previously-specified fixed price.

Vesting contracts accomplish several goals. The first is to provide price certainty for electricity retailers for a significant fraction of their wholesale energy needs. The second is to provide revenue certainty to the new owner of the generating facility. With a forward contract the new owner of the generation unit in our example already has a revenue stream each hour equal to the contract price times 400 MWh. These two aspects of vesting contracts protect suppliers and loads from the volatility of short-term market prices, because they only receive or pay the short-term price for production or consumption beyond the contract quantity. Finally, the existence of this fixed-price forward contract obligation has the beneficial impacts on the competitiveness of the short-term energy market described above.[7]

The contributing factor in the dramatic increase in short-term electricity prices during the summer of 2000 in California is the fact that the three large retailers–Pacific Gas and Electric, Southern California Edison, and San Diego Gas and Electric–purchased virtually all of their energy and ancillary services requirements from the day-ahead, hour-ahead, and real-time markets. When the amount of imports available from the Pacific Northwest was substantially reduced as a result of

[7]The price of energy sold under a vesting contract can also be used by the seller, typically the goverment, to raise or lower the purchase price of a generation facility. For same forward contract quantity, a higher energy price in the vesting contract raises the purchase price of the facility.

reduced water availability during the late spring and summer of 2000, the fossil fuel suppliers in California found themselves facing the significantly less elastic residual demand curves for their output. This fact, documented in Wolak (2003a), made the unilateral profit-maximizing mark-up of price above the marginal cost of producing electricity substantially higher during the summer and autumn of 2000 than it had been during the previous two years of the market. Moreover, particularly during the latter part of the autumn of 2000, the price of natural gas increased substantially relative to the levels that existed during the early part of 2000 and the previous two years. Because the vast majority of hours of the year natural gas-fired units set the price in California, this natural gas price increase led to a higher value for the marginal cost of the highest cost unit operating in California. Assuming that suppliers still bid to set market prices that satisfied equation (1), this higher marginal cost during the latter part of the 2000 lead to higher electricity prices for the same values of the elasticity of the residual demand curve facing each of the five large suppliers in the California electricity market.

5.3. Active Participation of Final Demand in Wholesale Market

Consider an electricity market with no variation in demand and supply across all hours of the day. Under these circumstances, it would be possible to build enough generation capacity to ensure that all demand could be served at some fixed price. However, the reality of electricity consumption and generation unit and transmission network operation is that demand and supply vary over time, often in an unpredictable manner. There is always a risk that a generation unit or transmission line will fail or that electricity consumer will decide to increase or decrease their consumption. This implies that there is always some likelihood that available capacity will be insufficient to meet demand. The increasing capacity share of renewable energy sources such as wind, solar, and hydro because of ongoing efforts to reduce greenhouse gas emissions, further increases the likelihood of energy shortfalls. Electricity can only produced from these sources when the wind is blowing, the sun is shining, or water is available behind the turbine.

There are two ways of eliminating a supply shortfall, either price must be increased to choke off demand, or demand must be rationed. Rationing is clearly an extremely inefficient way to ensure that supply equals demand. Many consumers willing purchase electricity at the prevailing price are unable to do so. Moreover, as has been discovered by politicians in all countries where rationing has occurred, the backlash associated with rationing can be devastating to those in charge.

Moreover, the indirect costs of rationing on the level economic activity can be substantial. In particular, preparing for and dealing with rationing periods also leads to substantial losses in economic output.

A superior approach to dealing with a shortfall of available supply relative to the level of demand at the prevailing price is to allow the retail price to rise to the level necessary to cause a sufficient number of consumers to reduce their consumption to bring supply and demand back into balance. Consumers that pay the hourly price of electricity for their consumption are not fundamentally different from generation unit owners responding to hourly price signals from a system reliability perspective. Let D(p) equal the consumer's hourly demand for electricity as function of the hourly price of electricity. Define SN(p) = D(0) - D(p), where D(0) is the consumer's demand for electricity at an hourly price equal to zero. The function SN(p) is the consumer's true willingness supply curve for "negawatts," reductions in the amount of megawatts consumed. Because D(p) is a downward sloping function of p, SN(p) is an upward sloping function of p. A generator with a marginal cost curve equal to SN(p) has the ability to provide the same reliability benefits as this consumer. However, an electricity supplier has the incentive to maximize the profits it earns from selling electricity in the spot market given its marginal cost function. In contrast, an industrial or commercial consumer with a negawatt supply curve, SN(p), can be expected to bid a willingness to supply negawatts into the spot market to maximize the profits it earns from selling its final output, which implies demand-bidding to reduce the average price it pays for electricity. Consequently, even though a generator and consumer may have the same true willingness-to-supply curve, each of them will use curve to pursue different goals. The supplier is likely to use it to exercise unilateral market power and raise market prices and the consumer is likely to use it exercise unilateral market power to reduce the price it pays for electricity. Wolak (2001) describes how a load-serving entity with some consumers facing the hourly wholesale price or a large consumer facing the hourly price could exercise market power on the demand side to reduce the average price it pays for a fixed quantity of electricity.

Besides allowing the system operator more flexibility in managing demand and supply imbalances, the presence of some consumers that alter their consumption in response to the hourly wholesale price also significantly benefits the competitiveness of the spot market. Figure 11 illustrates this point. The two residual demand curves are computed for the same value of SO(p).

For one, Q_D, is perfectly inelastic. For the other, $Q_D(p)$, is price elastic. As shown in the diagram, the slope of the resulting residual demand curve using $Q_D(p)$ is always flatter than the slope of the residual demand curve using Q_D. Following the logic used for the case of forward contracts, it can be demonstrated that for the same price and same value of residual demand, the elasticity of the residual demand curve using $Q_D(p)$, is always greater than the one using Q_D, because the slope of the one using $Q_D(p)$ is equal to $DR'(p) = Q_D'(p) - SO'(p)$, which is larger in absolute value than $-SO'(p)$, the slope of the residual demand curve using Q_D. Consequently, the competitive benefit of having final consumers pay the hourly wholesale price is that all suppliers will face more elastic residual demand curves, which will cause them to bid more aggressively into the spot market.

Politicians and policymakers often express the concern that the subjecting consumers to real-time price risk will introduce too much volatility into their monthly bill. These concerns are, for the most part, unfounded as well as misplaced. Borenstein (2005) suggests a scheme for facing a consumer with the hourly wholesale price for her consumption above or below a pre-determined load shape so that the consumer faces a monthly average price risk similar to a peak/off-peak time-of-use tariff.

It is important emphasize that some entity must manage short-term wholesale price risk and the risk of supply shortfalls. If a state regulator sets a fixed retail price or fixed pattern of retail prices throughout the day (time-of-use prices), some entity must still ensure that the over the course of the month or year, the retailer's total revenues less its transmission, distribution and retailing costs, must cover its total wholesale energy costs. If the regulator sets this fixed price too low relative to the current wholesale price then either the retailer or the government must pay the difference. Eventually, the government must make up the difference because it has the ability to impose taxes to fund its expenditures. However, these tax revenues are ultimately collected from consumers of electricity.

This is precisely the lesson learned by the citizens of California. When average wholesale prices rose above the average wholesale price implicit in the frozen retail price California consumers paid for electricity, retailers initially made up the difference. Eventually, these companies threatened to declare bankruptcy, in the case of Southern California Edison and San Diego Gas and Electric, and declared bankruptcy, in the case of Pacific Gas and Electric, so that the state of California took over purchasing wholesale power at even higher prices. The option to purchase all

electricity demand at a fixed-price or fixed-pattern of prices that does not vary with hourly system conditions is extremely valuable to consumers and extremely costly to the government.

This is nothing more than a re-statement of a standard prediction from the theory of stock options that the value of a call option on a stock is increasing in the volatility of the underlying security. However, different from the case of a call option on a stock, the fact that all California consumers had this option available to them and were completely shielded from any spot price risk in their electricity purchases (but not in their tax payments) made wholesale prices more volatile. By the logic of Figure 11, all suppliers faced a less elastic residual demand because all customers paid for their hourly electricity consumption at same fixed price or pattern prices rather than at the actual hourly real-time price. Therefore suppliers had a greater ability to exercise unilateral market, which led to higher average prices and greater price volatility.

By charging final consumers the same default price as generation units owners, final consumers will have strong incentive to become active participants in the wholesale market or purchase the appropriate short-term price hedging instruments retailers to eliminate their exposure to short-term price risk. These purchases of short-term price hedging instruments by final consumers increases the retailer's demand for fixed-price forward contracts from generation unit owners, which reduces the amount of energy that is actually sold at the short-term wholesale price.

Perhaps the most important, but most often ignored, lesson from electricity re-structuring processes in industrialized countries is the necessity of treating load and generation symmetrically. Symmetric treatment of load and generation means that unless a retail consumer signs a forward contract with an electricity retailer the default wholesale price he pays for all of his consumption is the hourly wholesale price. This is precisely the same risk that generation unit owners face. Unless it has signed a fixed-price forward contract with a load-serving entity or some other market participant, the price it receives for any short-term energy sales is the hourly short-term price. Just as very few suppliers are willing to risk selling all of their output in the short-term market, I would expect consumers to have similar preferences against too much reliance on the short-term market and would therefore be willing to sign a long-term contract for a large fraction of their expected hourly consumption during each hour of the month. Consistent with Borenstein's (2005b) logic, a residential consumer might purchase a right to buy a fixed load shape for each day at a fixed price for the next 12 months. This consumer would then be able to sell energy it does not consume during

any hour at the hourly wholesale price or purchase any power it needs beyond this baseline level at that same price. This type of pricing arrangement would result in a significantly less volatile monthly electricity bill than if the consumer made all of his purchases at the hourly wholesale price. If all customers purchased according to this sort of pricing plan then there would be no residual short-term price risk that the government needs to manage using tax revenues. All consumers manage the risk of high wholesale prices and supply shortfalls, according to their preferences for taking on short-term price risk. Moreover, because all consumers have an incentive to reduce their consumption during high-priced periods, wholesale prices are likely to be significantly less volatile. Rather than continuing to consume when wholesale prices rise, they now see this very high short-term price as the opportunity cost of consuming electricity for all of their consumption, with the important difference that if they consume less than their forward contract quantity, they are paid this very high price for each KWh they do not consume below that level.

Symmetric treatment of load and generation does not mean that a consumer is prohibited from purchasing a fixed-price full requirements contract for all of the electricity they might consume in a month, only that the consumer must pay the full cost of the retailer supplying this product. Imagine a gasoline retailer making a promise to its customers that they can purchase as much gasoline as they would like at a fixed price for an entire year. Given the volatility in wholesale gasoline prices, the price premium that a retailer would require to offer this service is likely to be very high. This sort of price premium should also exist for full requirements fixed-price contracts for electricity because the retailer is being asked to provide a fixed-price for any level of consumption by the final consumer.

The major technological roadblock to symmetric treatment of load and generation is the necessary metering technology to allow consumption to be measured on an hourly versus monthly basis. Virtually all existing meters at the residential level and the vast majority at the commercial and industrial level can only record total monthly consumption. Monthly meter reading means it is only possible to determine the total amount of KWh consumed between two consequence meter readings–the difference between the value on the meter at the start of the month and value at the beginning of the month is the amount consumed within the month. Without the metering technology necessary to record consumption for each hour of the month, it is impossible to determine precisely

how much customer consumed during each hour of the month, which is a necessary condition for symmetric treatment and load and generation.

The economic barriers to universal hourly metering have fallen over time. The primary cost associated with universal interval metering is the up-front cost of installing the system, although there is also a small monthly operating and maintenance cost. Wolak (2007a) describes the many technologies available. Many jurisdictions around the world have invested in universal interval meters for all customers and many others are in the process of doing so. For example, the three large retailers in California have been ordered by the CPUC to implement universal interval metering as a regulated distribution network service. The economic case for interval metering is primarily based on the cost savings associated with reading conventional meters. These automated interval meter systems eliminate the need for staff of the electricity retailer to visit the customer's premises to read the meter each month. Particularly in industrialized countries, where labor is relatively expensive, these savings in labor costs cover a significant fraction of the estimated cost of the automated meter reading system.

Wolak (2001) presents evidence for California that suggests that a portion of these costs would also be paid for by the lower wholesale electricity prices that result from the more competitive wholesale market that results from symmetric treatment of load and generation. The increased participation of negawatt suppliers in the wholesale market would face suppliers with more elastic residual demand curves which would cause them to submit willingness-to-supply curves closer to their marginal cost curves.

5.4. Economic Reliability versus Engineering Reliability of a Transmission Network

The presence of a wholesale market changes the definition of what constitutes a reliable transmission network. In order for it to be expected profit maximizing for generation unit owners to submit a bid curve close to their marginal cost curve, they must expect to face sufficiently elastic residual demand curves. For this to be the case, there must be enough transmission capacity into the area served by a generation unit owner so that any attempts to raise local prices will result in a large enough quantity of lost sales to make this bidding strategy unprofitable.

I define an economically reliable transmission network as one with sufficient capacity so that each location in the network faces sufficient competition from distant generation to cause local generation unit owners to compete with distant generators rather than cause congestion to a create

local monopoly market. In the former vertically-integrated utility regime, transmission expansions were undertaken to ensure the engineering reliability of the transmission network. A transmission network was deemed to be reliable from an engineering perspective if the vertically-integrated utility that controlled all of the generation units in the control area could maintain a reliable electricity supply to consumers despite unexpected generation and transmission outages.

The value of increasing the transmission capacity between two points still depends on the extent to which this expansion allows the substitution of cheap generation in one area for expensive generation in the other area. Under the vertically integrated monopoly regime, all differences across regions in wholesale energy payments were due to differences in the locational costs of production for the geographic monopolist. However, in the wholesale market regime, the extent of market power that can be exercised by firms at each location in the network can lead to much larger differences in payments for wholesale electricity across these regions.

Even if the difference in the variable cost of the highest cost units operating in two regions is less than $15/MWh, because firms in one area are able to exercise local market power, differences in the wholesale prices that consumers must pay across the two regions can be as high as the price cap on the real-time price of energy. For example, during early 2000 in the California market when the price cap on the ISO's real-time market was $750/MWh, because of congestion between Southern California (the SP15 zone) and Northern California (the NP15 zone), prices in the two zones differed by as much as $700/MWh, despite the fact that the difference in the variable costs of the highest cost units operating in the two zones was less than $15/MWh.

This example demonstrates that a major source of benefits from transmission capacity in a wholesale market regime is that it limits the ability of generation unit owners to use transmission congestion to limit the number of competitors they face. More transmission capacity into a local area implies that local generating unit owners face more competition from distant generation for a larger fraction of their capacity. Because these firms now face more competition from distant generation, they must bid more aggressively (supply curve closer to their marginal cost curve) over a wider range of local demand realizations to sell the same amount of energy they did before the transmission upgrade. In all cases, this more aggressive bidding brought about by the transmission upgrade will lower average wholesale energy prices on the congested side of the interface. Moreover, to the extent that the probability of congestion in one direction on an interface is

approximately equal to the probability of congestion in the opposite direction, the reduced opportunities for suppliers to exercise market power on both sides of the interface as a result of a transmission upgrade could reduce average wholesale prices at both locations.

The opportunity for generation unit owners to impact locational prices through their scheduling and bidding behavior creates another source of benefits of transmission upgrades in the wholesale market regime. In the vertically integrated monopoly regime, one rationale for upgrades of the monopolist's network was to manage the reliability risk associated with generation or transmission line outages. For example, an upgrade could be justified by the logic that if certain generating units became unavailable the supply shortfall could be temporarily served with distant, but more expensive, generating units. The reliability justification for such upgrades was that the cost of upgrade was less than the economic value created by the additional electricity that the consumers were able to consume because of the transmission upgrade.

Under the competitive market regime generators may have an additional incentive, besides that fact that unit is physically unable to operate, to declare their unit unavailable. They may find it profitable to create an artificial scarcity of generating capacity in a geographic area in order to increase the wholesale price they receive for the energy they do sell. This incentive to withhold generating capacity did not exist in the regulated monopoly regime. The monopolist was required by law to serve all load demanded at the regulated retail price. However, in the wholesale market regime, if a generator is able to raise the price it receives by 100 percent by withholding less than 10% of its capacity, it is likely to find this behavior profitable.

Consequently, in the wholesale market regime, reliability risk has an additional dimension because of the incentive for generation unit owners to withhold capacity from the market to increase prices if they do not face sufficient competition. For example, few, if any, market observers would have predicted as late as August 2000 that the California ISO would experience a daily average of approximately 10,000 MW of generating units off-line during the eight-month period November 2000 to May of 2001. Additional transmission capacity can render physical withholding strategies, which may lead to load curtailments, less profitable and therefore less likely to occur.

Understanding how transmission upgrades can increase the elasticity of the residual demand a supplier faces requires only a slight modification of the discussion surrounding Figure 10. Suppose that 9,500 MWh of demand is all located on the other side of a transmission line with 9,000 MW of

capacity and the supplier under consideration owns 1000 MW of generation local to the demand. Suppose there is 12 firms each of which own 1,000 MW of capacity located on the other side of the interface. In this case, the local supplier is pivotal for 500 MWh of energy because local demand is 9,500 MWh but only 9,000 MWh of energy can get into the local area because of transmission constraints. Note that there is 12,000 MW of generation capacity available to serve the local demand. It just can't get into the region because of transmission constraints. We can now re-interpret $SO_1(p)$ in Figure 10 as the aggregate bid supply curve of the 12 firms competing to sell energy into the 9,000 MW transmission line.

Suppose the transmission line is now upgraded to 9,500 MW. From the perspective of the local firm this results in $SO_2(p)$ to serve the local demand, which means that the local supplier is no longer pivotal. Before the upgrade the local supplier faced the residual demand curve $DR_1(p)$ in Figure 10 and after the upgrade it faces $DR_2(p)$, which is more elastic than $DR_1(p)$ at all price levels. This is the mechanism by which transmission upgrades increases the residual demand electricity a supplier faces and the overall competitiveness of the wholesale electricity market.

The California Independent System Operator's (ISO) Transmission Expansion Assessment Methodology (TEAM) incorporates the increased wholesale competition benefits of associated transmission expansions. Awad et. al. (2007) presents the details of this methodology and apply it to a proposed transmission expansion from Arizona into Southern California–the Palo Verde-Devers Line No. 2 upgrade. The authors find that the result of increased competition that generation unit owners in California face from generation unit owners located in Arizona is a major source of benefits from the upgrade. These benefits are much larger for system conditions with low levels of hydroelectric energy available from the Pacific Northwest and very high natural gas prices, because this transmission expansion allows more electricity imports more from the Southwest, where the vast majority of electricity is produced using coal.

6. Role of Regulatory Oversight in Market Design Process

Regulatory oversight of the wholesale market regime is perhaps the most difficult aspect of the market design process. The regulatory process in the vertically-integrated regime focuses on setting just and reasonable prices through an administrative procedure that determines the firm's prudently incurred costs and sets a price that allows the firm the opportunity to recover these costs. The regulatory process in the wholesale market regime focuses on the far more difficult and risky

task of setting market rules that will yield, through the unilateral-optimizing actions of market participants, just and reasonable prices to final consumers. The rules that govern the operation of the generation, transmission, distribution and retailing sectors of the industry all impact the retail prices paid by final consumers. As Section 4 makes clear, regulatory oversight of the wholesale market regime is a considerably more difficult because of the individual rationality constraint that each firm will choose its actions to influence the revenues it receives and costs it incurs to maximize its objective function subject to these market rules.

Regulatory oversight is further complicated by the fact that actions taken by the regulator to correct a problem in one aspect of the wholesale market can impact the individual rationality constraint faced by other market participants. The change in behavior by these market participants can lead to market outcomes that create more adverse economic consequences than the problem that caused the regulator to take action in the first place. This logic implies that the regulator must examine the full implications of any proposed market rule changes or other regulatory interventions, because once they have been implemented market participants will alter the constraint set they face and maximize their objective function subject to this new constraint set, consistent with their individual rationality constraint.

Despite the significant challenges faced by the regulatory process in the wholesale market regime, the restructured electricity supply industries that have ultimately delivered the most benefits to electricity consumers are those with a credible and effective regulatory process. This section summarizes the major tasks of the regulatory process in the wholesale market regime. The first is to provide what I call "smart sunshine regulation." This means that the regulatory process gathers a comprehensive set of information about market outcomes, analyzes it, and make it available to the public in a manner and form that ensures compliance with all market rules and allows the regulatory and political process to detect and correct market design flaws in a timely manner. Smart sunshine regulation is the foundation for all of the tasks the regulatory process must undertake in the wholeale market regime.

For the reasons discussed in Section 5.4, the regulatory process must also take a more active role in managing the configuration of the transmission network than it did in the former vertically-integrated regime. Because the real-time wholesale market operator is a monopoly provider of this service, the regulator must monitor its performance. The regulatory process must also oversee the

performance of the retailing and energy trading sectors. Finally, the regulatory process must have the ability to take actions to prevent significant the wealth transfers and deadweight losses that can result from the legal (under US antitrust law) exercise of unilateral market power in wholesale electricity markets. This is perhaps the most challenging task the regulatory process faces because knowledge that the regulator will take actions to prevents these transfers and deadweight losses can limit the incentive market participants have to take costly actions to prevent the exercise of unilateral market power.

6.1. Smart Sunshine Regulation

A minimal requirement of any regulatory process is to provide "intelligent sunshine" regulation. The fundamental goal of regulation is to cause a firm to take actions desired by the regulator that it would not otherwise do without regulatory oversight. For example, without regulatory oversight, a vertically-integrated monopoly is likely prefer to raise prices to some customers and/or refuse to serve others. One way to cause a firm to take actions desired by the regulator is to use the threat of an unfavorable public reaction to discipline the behavior of the firm. In the above example, if the firm is required by law to serve all customers at the regulated price, a straightforward way to increase likelihood that the firm complies is for the regulator to disclose to the public instances when the firm denies service to a customer or charges a higher or lower price.

In order provide effective smart sunshine regulation, the regulator must have access to all information needed to operate the market and be able to perform analyses of this data and release the results to the public. At the most basic level, the regulator should be able to replicate market-clearing prices and quantities given the bids submitted by market participants, total demand, and other information about system conditions. This is necessary for the regulator to verify that the market is operated in a manner consistent with what is written in the market rules.

A second aspect of "smart sunshine regulation" is public data release. There are market efficiency benefits to public release of all data submitted to the real-time market and produced by the system operator. As discussed in Section 4.2, if a very small fraction of energy sales takes place at the real-time price, this limits the incentive for large suppliers to exercise unilateral market power in the short-term wholesale market. With adequate hedging of short-term price risk by electricity retailers, this market is primarily an imbalance market operated primarily for reliability reasons, where retailers and suppliers buy and sell small amounts of energy to manage deviations between

their forward market commitments and real-time production and consumption. Because all market participants have a common interest in the reliability of the transmission network, immediate data release serves these reliability needs.

Wholesale markets that currently exist around the world differ considerably in terms of amount of data they make publicly available and the lag between the date the data is created and the date it is released to the public. Nevertheless, among industrialized countries there appears to be a positive correlation between the extent to which data submitted or produced by the system operator is made publicly available and how well the wholesale market operates. For example, the Australian electricity market makes all data on bids and unit-level dispatch publicly available the next day. Australia's National Electricity Market Management Company (NEMMCO) posts this information by market participant name on its website. The Australian electricity market is generally acknowledged to be one of the best performing re-structured electricity markets in the world (Wolak, 1999).

The former England and Wales electricity pool kept all of the unit-level bid and production data confidential. Only members of the pool could gain access to this data. It was generally acknowledged to be subject to the exercise of substantial unilateral market power by the larger suppliers, as documented by Wolak and Patrick (1997) and Wolfram (1999). The UK government's displeasure with pool prices eventually led to the New Electricity Trading Arrangements (NETA) which began operation on March 27, 2001. Although these facts do not provide definitive proof that rapid and complete data release enhances market efficiency, the best available information on this issue provides no evidence that withholding this data from the public scrutiny enhances market efficiency.

The sunshine regulation value of public data release is increased if the identity of the market participant and the specific generation unit associated with each bid, generation schedule, or output level is also made public. Masking the identity of the entity associated with a bid, generation schedule or output level, as is done in all US wholesale markets, limits the ability of the regulator to use the threat of adverse public opinion to discipline market participant behavior. Under a system of masked data release, market participants can always deny that their bids or energy schedules are the ones exhibiting the unusual behavior. The primary value of public data release is putting all market participants at risk for explaining to the public that their actions are not violation of the intent

of the wholesale market rules. In all US markets, the very long lag between the date the data is produced and the date it is released to the public, at least six months, and the fact that the data is released without identifying the specific market participants, eliminates much of the smart sunshine regulation benefit of public data release.

Putting market participants at risk for explaining their behavior to the public is different from requiring them to behave in a manner that it is inconsistent with their unilateral profit-maximizing interests. A number of markets have considered implementing "good behavior conditions" on market participants. The most well-known attempt was the United Kingdom's (UK) consideration of a Market Abuse License Condition (MALC) as a pre-condition for participating in its wholesale electricity market. The fundamental conflict raised by these "good behavior" clauses is that they can prohibit behavior that is in the unilateral profit-maximizing interests of a supplier that is also in the interests of consumers. These "good behavior" clauses do not correct the underlying market design flaw or implement a change in the market structure to address the underlying cause of the harm from the unilateral exercise of market power. They simply ask that the firm be a "good citizen" and not maximize profits. In testimony to the United Kingdom Competition Commission, Wolak (2000b) made these and a number of other arguments against the MALC, which the Commission eventually decided against implementing.

Another potential benefit associated with public data release is that it enables independent third-parties to undertake analyses of market performance. The US policies on data release limit the benefits from this aspect of a public data release policy. Releasing data with the identities of the market participant masked makes it impossible to definitively match data from other sources to specific market participants. Virtually all market performance measures require matching data on unit-level heat rates or input fuel prices obtained from other sources to specific generation units. Strictly speaking, this is impossible to do if the unit name or market participant name is not matched with the generation unit.

A long time lag between the date the data is produced and the date it is released also greatly limits the range of questions that can be addressed with this data and regulatory problems that it can address. Taking the example of the California electricity crisis, by January 1, 2001, the date that masked data from June of 2000 was first made available to the public (because of a six-month data release lag), the exercise of unilateral market power in California had already resulted in more than

$5 billion in overpayments to suppliers in the California electricity market as measured by BBW (2002). Consequently, a long time lag between the date the data is produced and the date it is released to the public has an enormous potential cost to consumers that should be balanced against the benefits of delaying the data release.

The usual argument against immediate data release is that suppliers could use this information to coordinate their actions to raise market prices through sophisticated tacit collusion schemes. However, there are a number of reasons why these concerns are much less relevant for the release of data from a short-term bid-based wholesale market. First, as discussed above, in a wholesale electricity market with the levels of hedging of short-term price risk necessary to leave large suppliers with little incentive to exercise unilateral market power in the short-term market, very little energy is actually sold at the short-term price. The short-term market is primarily a venue for buying and selling energy imbalances. With adequate levels of hedging of short-term price risk, both suppliers and retailers would rarely have significant positions on either side of the short-term market. Therefore, they would have little incentive to raise prices in short-term market through their unilateral actions or through coordinated behavior.

Nevertheless, without adequate levels of hedging of short-term price risk, the immediate availability of information on bids, schedules and actual unit-level production could allow suppliers to design more complex state-dependent strategies for enforcing collusive market outcomes. However, it is important to bear in mind that coordinated actions to raise market prices are illegal under US antitrust law and under the competition law in virtually all countries around the world. The immediate availability of this data means that the public also has access to this information and can undertake studies examining the extent to which market prices differ from the competitive benchmark levels described in BBW (2002). Keeping this data confidential or releasing it only after a long time lag prevents this potentially important form of public scrutiny of market performance from occurring.

In contrast to data associated with the operation of the short-term wholesale market, releasing information on forward market positions or transactions prices for specific market participants is likely to enhance the ability and incentive of suppliers to raise the prices retailers pay for these hedging instruments. Large volumes of energy are likely to be traded in this market and suppliers typically sell these products and retailers and large customer typically buy these products. Forward

market position information about a market participant is unnecessary to operate the short-term market, so there is little reliability justification releasing this data.

There is a strong argument for keeping any forward contract positions the regulator might collect confidential. As noted in Wolak (2000a), the financial forward contract holdings of a supplier are major determinants of the aggressiveness of its bids into the short-term market. Only if a supplier is confident that it will produce more than its forward contract obligations will it have an incentive to bid or schedule its units to raise the market price. Suppliers recognize this incentive created by forward contracts when they bid against competitors with forward contract holdings. Consequently, public disclosure of the forward contract holdings of market participants can convey useful information about the incentives of individual suppliers to raise market prices, with no countervailing reliability or market-efficiency enhancing benefits.

A final aspect of the data collection portion of the regulatory process is concerned with scheduled outage coordination and forced outage declarations. A major lesson from wholesale electricity markets around the world is the impossibility of determining whether a unit that is declared out-of-service can actually operate. Different from the former vertically integrated regime, declaring a "sick day" for a generation unit–saying that it is unable to operate when in reality it could safely operate–can be a very profitable way for a supplier to withhold capacity from the market in order to raise the wholesale price. To limit the ability of suppliers to use their planned and unplanned outage declarations in this manner, the market operator and regulator must specify clear rules for determining a unit's planned outage schedule and for determining when a unit is forced out.

To limit the incentive for "sick day" unplanned generation outages, the system operator could specify the following scheme for outage reporting. Unless a unit is declared available to operate up to its full capacity, the unit is declared fully out or partially out depending on the amount capacity from the unit bid into the market at any price at or below the current price cap. This definition of a forced outage eliminates the problem of determining whether a unit that does not bid into the market is actually able to operate. A simple rule is to assume the unit is be forced out because the owner is not offering this capacity to the market. The system operator would therefore only count capacity from a unit bid in at a price at or below the price cap as available capacity. Information on unit-level forced outages according to this definition could then be publicly disclosed each day on the system operator's web-site.

This disclosure process cannot prevent a supplier from declaring a "sick day" to raise the price it receives for energy or operating reserves that it sells from other units it owns. However, the process can make it more costly for the market participant to engage in this behavior by registering all hours when capacity from a unit is not bid into the market as forced outage hours. For example, if a 100 MW generation unit is neither bid nor scheduled in the short-term market during an hour, then it is deemed to be forced out for that hour. If this unit only bids 40 MW of the 100 MW at or below the price or bid cap during an hour, then the remaining 60 MW is deemed to be forced out for than hour. The regulator can then periodically report forced outage rates based on this methodology and compare these outage rates to historical figures from these units before re-structuring or from comparable units from different wholesale markets. The regulator could then subject the supplier to greater public scrutiny and adverse publicity for significant deviations of the forced outage rates of its units relative to those from comparable units.

A final issue associated with smart sunshine regulation is ensuring compliance with market rules. The threat of public scrutiny and adverse publicity is the regulator's first line of defense against market rule violations. However, an argument, based on the logic of the individual rationality constraint implies that the regulator must make the penalties associated with any market rule violations more than the benefits that the market participant receives from violating that market rule. Otherwise market participant may will find it unilaterally profit-maximizing to violate the market rules. One lesson from the activities of many firms in the California market and other markets in the US is that if the cost of a market rule violation is less than the financial benefit the firm receives from violating the market rule, the firm will violate the market rule and pay the associated penalties as a cost of doing business.

6.2. Detecting and Correcting Market Design Flaws

Bid-based wholesale electricity markets can have market design flaws that have little impact on market outcomes during most system conditions but result in large wealth transfers under certain system conditions. Consequently, an important role of the regulatory process is to detect and correct market design flaws before circumstances arise that cause them to produce large wealth transfers and significant deadweight losses.

The experience of the California market illustrates this point. From its start in April 1998 until April 2000, the California market set prices that were very close to those that would occur if

no suppliers exercised unilateral market power, what Borenstein, Bushnell and Wolak (2002), hereafter BBW, call the competitive benchmark price. BBW compute this competitive benchmark price using daily data on input prices and the technical operating characteristics of all generation units in California and the hourly willingness to supply importers to construct a counterfactual competitive supply curve that they intersect with the hourly market demand. During the first two years of the California market, the average difference between the actual hourly market price and the hourly competitive benchmark price computed using the BBW methodology is less than or very close to equal to those computed by Mansur (2003) for the PJM market and Bushnell and Saravia (2003) for the New England market using this same methodology. Actual market prices very close to competitive benchmark prices occurred in spite of the fact that virtually all of the wholesale energy purchases by the three large California retailers were made through the day-ahead or real-time market.

This over-reliance on short-term markets led to actual prices that were not substantially different from competitive benchmark prices because there was plenty of hydroelectric energy in California and the Pacific Northwest and low cost fossil-fuel energy from the Southwest during the summers of 1998 and 1999. Any attempts by fossil fuel suppliers in California to withhold output to raise short-term prices were met with additional supply from these sources with little impact on market prices. In the language Section 5, these fossil fuel suppliers faced very elastic residual demand curves because of the flat willingness-to-supply functions offered by hydroelectric suppliers and importers. Given these system conditions, California's fossil fuel suppliers found it unilaterally profit-maximizing to offer each of their generation units into the day-ahead and real-time markets at very close to the marginal cost of production.

These unilateral incentives changed in the summer of 2000 when the amount of hydroelectric energy available from the Pacific Northwest and Southwest was significantly less than was available during the previous two summers. Wolak (2003a) shows that this event led the five largest fossil fuel electricity suppliers in California to face significantly less elastic residual demand curves because of the less aggressive supply responses from importers to California than they did during the first two summers of the wholesale market. As a consequence, these suppliers found it in their unilateral interest to exploit these less elastic residual demand curves and withhold output from the short-term market in order to raise wholesale electricity prices in California. During the summer

months of June to September of 2000, the average difference between the actual price and the BBW competitive benchmark price was more than $70/MWh, which is more than twice the average price of wholesale electricity during the first two years of the market of $34/MWh.

The California experience demonstrates that some market design flaws, in this case insufficient hedging of short-term price risk by electricity retailers, can be relatively benign under a range of system conditions. However, when system conditions conducive to the exercise of unilateral market power occur, this market design flaw can result in substantial wealth transfers from consumers to producers and economically significant deadweight losses. BBW (2002) present estimates of these magnitudes for the period June 1998 to October 2000.

It important to emphasize that these wealth transfers appear to have occurred without co-ordinated actions among market participants that violated US antitrust law. Despite extensive multi-year investigations by almost every state-level antitrust and regulatory commission in the western US, the US Department of Justice Antitrust Division, the Federal Energy Regulatory Commission, and numerous Congressional committees, no significant evidence of coordinated actions to raise wholesale electricity prices in the WECC during the period June 2000 to June 2001 has been uncovered. This outcome occurred because US antitrust law does not prohibit firms from fully exploiting their unilateral market power. This fact emphasizes the need, discussed later in this section, for the regulator to have the ability to intervene when the unilateral exercise of market power is likely to result in significant wealth transfers.

Identifying and correcting market design flaws requires a detailed knowledge of the market rules and their impact on market outcomes. This aspect of the regulatory process heavily relies on the availability of the short-term market outcome data and other information collected by the regulator to undertake smart sunshine regulation. Another important role for smart sunshine regulation is to analyze market outcomes to determine which market rules might be enhancing the ability of suppliers to exercise unilateral market power or increasing the likelihood that the attempts of suppliers to coordinate to raise market prices will be successful.

6.3. Oversight of Transmission Network and System Operation

There are also important market competitiveness benefits from regulatory oversight of the terms of conditions for new generation units to interconnect to the transmission network and determine whether transmission upgrades should take place and where they should take place. As

discussed in Wolak (2003c), in the wholesale market regime transmission capacity has an additional role as a facilitator of commerce. As noted in Section 5.4, expansion of the transmission network typically increases the number of independent wholesale electricity suppliers that are able to compete to supply electricity at locations in the transmission network served by the upgrade, which increases the elasticity of the residual demand curve faced by all suppliers at those locations. An industry-specific regulator armed with the data and experienced with monitoring market performance is well-suited to develop the expertise necessary to determine the transmission network that maximizes the competitiveness of the wholesale electricity market.

The Independent System Operator (ISO) that operates the real-time market is a new entity requiring regulatory oversight in the wholesale market regime. The system operation function was formerly part of the vertically-integrated utility. Because a wholesale market provides open-access to the transmission network under equal terms and conditions to all electricity suppliers and retailers, an independent entity is needed to operate the transmission network to maintain system balance in real-time. The ISO is the monopoly supplier of real-time market and system operation services and for that reason independent regulatory oversight is needed to ensure that it is operating the grid in as close as possible to a least-cost manner to benefit market participants rather than the management and staff of the ISO.

In virtually all markets in the US, the day-ahead forward energy and generation reserves markets are operated by the ISO. In this case, the ISO is also a monopoly supplier of day-ahead market services, which creates addition responsibilities for the regulatory process. In the US, integration of the day-ahead market with real-time system operation is justified based on the fact that many generation units have long-start times, so there are potential reliability consequences associated with the ISO not operating a day-ahead forward market. In a number of other countries of the world, the ISO does not operate a formal day-ahead forward market. Instead, there are competing day-ahead forward markets offered by third-parties, so that less regulatory oversight of these forward markets is necessary.

A final issue with respect to regulatory oversight of the transmission network and system operation function is the fact that the ISO has substantial expertise with operating transmission network. Consequently, the regulator may find it beneficial to allow the ISO to play a leading role in process of determining competition expansions to the transmission network.

6.4. Oversight of Trading and Retailing Sectors

Traders and competitive retailers are the final class of new market participants requiring regulatory oversight. Traders typically buy something they have no intention of consuming and sell something they do not or cannot produce. In this sense, energy traders are no different from derivative securities traders who buy and sell puts, calls, swaps and futures contracts. Traders typically take bets on the direction that electricity prices are likely to move between the time the derivative contract is signed and the expiration date of the contract. Securities traders profit from buying a security at a low price and selling it later for a higher price, or selling the security at a high price and buying it back later at a lower price. Energy traders can also serve a risk management role by taking on risk that other market participants would prefer not to bear.

Competitive retailers are specific type of energy trader. They provide short-term price hedging services for final consumers to compete with the products offered by the incumbent retailer. They purchase and sell hedging instruments with the goal of providing a retail electricity at prices final consumers find attractive. The major regulatory oversight challenge for the competitive retailing sector is to ensure that retailers do not engage in excessive risk-taking. For example, a retailer could agree to sell electricity to final consumers at a low fixed retail price by purchasing the necessary electricity from the short-term wholesale market. However, if short-term wholesale prices rise, this retailer might then be forced into bankruptcy because of its fixed-price commitment to sell electricity to final consumers at a price that does not recover the cost of the wholesale electricity. The regulatory process must ensure that retailers adequately hedge any fixed-price forward market commitments they provide to final consumers.

A trader activity that has created considerable controversy among politicians and the press is attempts to exploit potential price differences for the same product across time or locations. For the case of electricity, this could involve exploiting the difference between the day-ahead forward price for electricity for one hour of the day and the real-time price of electricity for that same hour. Locationally, this involves buying the right to inject electricity at one node and selling the right to inject electricity at another node. This is often incorrectly described as buying electricity at one node and selling it at another node. As the discussion surrounding Figure 7 demonstrates, it is not possible to take possession of electricity and transport it from one node to another. Consequently, selling a 1 MWh injection of electricity at node A and buying a 1 MWh withdrawal at node B in the

day-ahead market is taking a gamble on the difference in the direction and magnitude of congestion between these two locations in the transmission network. In the real-time market the trader can fulfill his obligation to inject at node A by purchasing electricity at the real-time price at node A and his obligation to withdraw at node B by selling energy at the real-time price at node B. In this case, the trader neither produces nor consumes electricity in real-time, but its profit on these transactions is the difference between the day-ahead prices at nodes A and B less the difference of the real-time prices at nodes B and A.

Virtually all of these transactions involve a significant risk that the trader will lose money. For example, if a trader sells 1 MWh at the day-ahead price at node A and the real-time price turns out to be higher than day-ahead price at node A, then the trader must fulfill the commitment to provide 1 MWh at node A by purchasing at the higher real-time price. This transaction earns the trader a loss equal to the difference between the real-time and day-ahead prices.

Advocates of energy trading often speak of traders providing "liquidity" to a market. A liquid market is one where large volumes can be bought or sold without causing significant market price movements. Viewed from this perspective, traders can benefit market efficiency. However, there may be instances when the actions of traders degrade market efficiency, by exploiting market design flaws. As Wolak (2003b) notes, virtually all of the Enron trading strategies described in the three memos released by FERC in the Spring of 2002 could be classified as risky trading strategies that had the potential to enhance market efficiency. Only a few clearly appeared to degrade system reliability or market efficiency. Consequently, a final challenge for the regulatory process in the wholesale market regime is to ensure that the profit-maximizing activities of traders enhance, rather than detract, from market efficiency.

6.5. Protecting Against Behavior Harmful to Market Efficiency and System Reliability

The final responsibility for the regulator is to deter behavior that is harmful to system reliability and market efficiency that occurs despite public disclosure of data and market participant behavior and penalties for publicly-observed, objective market rule violations. This is the most complex aspect of the regulatory process to implement, but it also has the potential to yield the greatest benefit. It involves a number of inter-related tasks. In a bid-based market, the regulator must design and implement a local market power mitigation mechanism, which is the most frequently invoked example of an intervention into the market to prevent behavior harmful to market

efficiency and system reliability. In general, the regulator must determine when any type of market outcome causes enough harm to some market participants to merit explicit regulatory intervention. Finally, if the market outcomes become too harmful, the regulator must have the ability to temporarily suspend market operations. All of these tasks require a substantial amount of subjective judgement on the part of the regulatory process.

In all bid-based electricity markets a local market power mitigation mechanism is necessary to limit the bids a supplier submits when it faces is insufficient competition to serve a local energy need because of combination of the configuration of the transmission network and concentration of ownership of generation units. An LMPM mechanism is a pre-specified administrative procedure (usually written into the market rules) that determines: (1) when a supplier has local market power worthy of mitigation, (2) what the mitigated supplier will be paid, and (3) how the amount the supplier is paid will impact the payments received by other market participants. Without a prospective market power mitigation mechanism system conditions are likely to arise in all wholesale markets when almost any supplier can exercise substantial unilateral market power. It is increasingly clear to regulators around the world, particularly those that operate markets with limited amounts of transmission capacity, that formal regulatory mechanisms are necessary to deal with the problem of insufficient competition to serve certain local energy needs.

The regulator is the first line of defense against harmful market outcomes. Persistent behavior by a market participant that is harmful to market efficiency or system reliability is typically subject to penalties and sanctions. In order to assess these penalties, the regulator must first determine intent on the part of the market participant. The goal of this provision is to establish a process for the regulator to intervene to prevent a market meltdown. As discussed in Wolak (2004), there are instances when actions very profitable to one or a small number of market participants can be extremely harmful to system reliability and market efficiency. A well-defined process must exist for the regulator to intervene to protect market participants and correct the market design flaw facilitating this harm. Wolak (2004) proposes such an administrative process for determining behavior harmful to system reliability and market efficiency that results from the exercise of unilateral market power by one or more market participants.

The regulator may also wish to have the ability to suspend market operations on a temporary basis when system conditions warrant it. The suspension of market operations is an extreme

regulatory response that requires a pre-specified administrative procedure has been followed and it has been determined that it is the only option available to the regulator to prevent significant harm to market efficiency and system reliability. As has been demonstrated in various countries around the world, electricity markets can sometimes become wildly dysfunctional and can lead to significant wealth transfers and deadweight losses over a very short period time. Under these sorts of circumstances, the regulator should have the ability to suspend market operations temporarily until the problem can be dealt with through a longer-term regulatory intervention or market rule change. Wolak (2004) proposes a process for making such a determination.

Different from the case of the vertically-integrated utility regime, the regulator must be forward-looking and fast-acting, because wholesale markets provide extremely high-powered incentives for firm behavior, so it does not take very long for a wholesale electricity market to produce enormous wealth transfers from consumers to producers and significant deadweight losses. The California electricity crisis is an example of this phenomenon. The Federal Energy Regulatory Commission (FERC), the entity that regulates wholesale markets in the US, waited almost six months from the time it first became clear that there was substantial unilateral market power exercised in the California market before it took action. As Wolak (2003b) notes, when FERC finally did take action in December 2000, it did so with little, if any, quantitative analysis of market performance, in direct contradiction of the fundamental need for smart sunshine regulation of the wholesale market. Wolak (2003b) argues that the actions FERC took at this time increased the rate at which wealth transfers occurred. Wolak, Nordhaus, and Shapiro (2000) discuss the likely impact, which as Wolak (2003b) notes, also turned out to be the eventual impact, of the FERC's December of 2000 action.

7. Common Market Design Flaws and Their Underlying Causes

This section describes a several common market design failures and uses the framework of Sections 4 to 6 to diagnose their underlying causes. These include excessive focus by the regulatory process on spot market design, inadequate divestiture of generation capacity by the incumbent firms, lack of an effective local market power mitigation mechanism, price caps and bid caps on short-term markets, and an inadequate retail market infrastructure.

7.1. Excessive Emphasis on Spot Market Design

Relative to other industrialized countries, the wholesale market design process in the US has focused much more on the details of short-term energy and operating reserves markets. This is design focus sharply contrasts with the focus of the restructuring processes in many developing countries, particularly in Latin America. These countries aim to foster an active forward market for energy and many of them impose regulatory mandates for minimum percentages of forward contract coverage of final demand at various time horizons to delivery. The short-term market is operated primarily to manage system imbalances in real-time, and in the majority of Latin American countries this process operates based on the ISO's estimate of the variable cost of operating each generation unit, not the unit owner's bids.

Joskow (1997) argues that the major source of benefits from electricity industry restructuring is likely to come from more efficient new generation investment decisions, rather than from more efficient operation of existing generation units to meet final demand. Nevertheless, there does appear to be evidence that individual generation units operating in a restructured wholesale market environment tend to be operated in a more efficient manner. Fabrizio, Rose and Wolfram (2007) use data on annual plant-level input data to compare the relative efficiency of municipally-owned plants versus those owned by investor-owned utilities in the pre- versus post-restructuring regimes. They find that the efficiency of municipally-owned units was largely impacted by restructuring, but those plants owned by investor-owned utilities in restructured state significantly reduced non-fuel operating expenses and employment. Bushnell and Wolfram (2005) use data on hourly fossil fuel use from the Environment Protection Agency's (EPA), Continuous Emissions Monitoring System (CEMS) to investigate changes in operating efficiency, the rate at which raw energy is translated into electricity, at generation units that have been divested from investor-owned utility to non-utility ownership. They find that fuel efficiency (or more precisely average heat rates) improved by about 2 percent following divestiture. They also find that non-divested plants that were subject to incentive regulation also realized similar magnitudes of average heat rate improvements. Unless the vast majority of final demand is covered by fixed-price forward contracts or other hedging arrangements that effectively fix the price received by suppliers to serve the vast majority of final demand, these operating efficiency gains are unlikely to be passed on to final consumers.

The magnitude of unilateral market power exercised in US electricity markets documented in the studies by BBW (2002), Joskow and Kahn (2002), Mansur (2003) and Bushnell and Saravia (2003), the magnitude of these operating efficiency gains are substantially smaller than the average percentage mark-up of market prices over estimated competitive benchmark prices. This implies that these operating efficiency gains are most likely being captured by generation unit owners rather than electricity consumers, unless there is adequate fixed-price hedging between suppliers and final demand.

This is fundamental problem with a perspective that emphasizes short-term market design. It is extremely difficult to establish a workably competitive short-term market under moderate to high demand conditions without a substantial amount of final demand covered by fixed-priced long-term contracts. A very unconcentrated generation ownership structure, far below the levels that currently exist in all US markets, would be necessary to achieve competitive markets outcomes under these demand conditions in the absence of high levels of fixed-price forward contract coverage of final demand. By the logic of Section 5.3, the greater is the share of total generation capacity owned by the largest firm in the market, the lower is the level of demand at which short-term market power problems are likely to show up, unless a substantial fraction of this larger supplier's expected output has been sold in a fixed-price forward contract. For virtually any number of suppliers and distribution of generation capacity ownership among these suppliers in a wholesale market without forward contracting, there is a level of demand at which significant spot market power problems will arise.

It is important to emphasize that having adequate generation capacity installed to serve demand according to the standards of the former vertically-integrated utility regime does very little to prevent the exercise of substantial unilateral market power in a wholesale market regime with inadequate fixed-price forward contracting. A simple example emphasizes this point. Suppose that there are five firms. One owns 300 MW of generation capacity, the second 200 MW, and the remaining three each own 100 MW, for a total of 800 MW. If demand is 650 MWh, then there is adequate generation capacity to serve demand, but it is extremely likely that spot prices will be at the bid cap, because the two largest suppliers know they are pivotal--some of their generation capacity is needed to meet demand regardless of the actions of their competitors. If all suppliers have zero fixed-price forward contract commitments to retailers, even at a demand slightly above

500 MW, the largest supplier is pivotal and therefore able to exercise substantial unilateral market power.

The presence of some price-responsive demand does not alter the basic logic of this example. For example, suppose that 100 MWh of the 650 MWh of demand is willing to respond to wholesale prices, then the demand can simply be treated as an additional 100 MW negawatt supplier in the calculation of what firms are pivotal at this level of demand. In this case, the firm than owns 300 MW of generation capacity would still be pivotal because after subtracting the capacity of all other firms besides this one, including the 100 MW of negwatts, from system demand, 50 MWs is needed from this supplier or total demand will not be met. Under this scenario, unless the largest supplier has fixed-price forward contract to supply of at least 50 MWh, consumers will be subject to substantial market power in the short-term energy market at this demand level.

One solution proposed to the problem of market power in short-term energy markets with insufficient forward contracting is to build additional generation capacity so that system condition never arise where suppliers have the ability to exercise unilateral market power in the spot market. In the above example of the five suppliers with no price responsive final demand and a total demand of 650 MWh, this would require constructing an additional 150 MW by new entrants or the four remaining smaller firms, with at least 50 MW being constructed by any entity but the first and second largest firms. This amount of new generation capacity distributed among new entrants and the remaining firms in the market would prevent any supplier from being pivotal in the short-term market with no forward contracting at a demand of 650 MWh.

There are several problems with this solution. First, it typically requires substantial excess capacity, particularly in markets where generation capacity ownership is concentrated. In the above example, there would now be at least 950 MW of generation capacity in the system to serve a demand of 650 MWh. Second, there is no guarantee this new generation capacity will be built by the entities necessary for the two largest firms not be pivotal. Finally, this excess capacity must be paid for or it will exit the industry. This excess capacity creates a set of stakeholders advocating for additional revenues to generation unit owners beyond those obtained from energy sales. Finally, this excess capacity is likely to depress short-term energy prices and dull the incentive for active demand-side participation in the wholesale energy market, which should lead to more calls for additional payments to generation owners to compensate for their energy market revenue shortfalls.

A far less costly solution to the problem of market power in short-term energy and reserve markets is for retailers to engage in fixed-priced forward contracts for a significant fraction of their final demand. This solution does not require installing additional generation capacity. In fact, it provide strong incentives for suppliers to construct the minimum amount generation capacity needed to meet these fixed-price, forward contract obligation for energy and reserves. To see the relationship between the level of fixed-price forward contract coverage of final demand and the level of demand at which market power problems arise in the short-term market, consider the same example except that all suppliers have sold 80 percent of their generation capacity in fixed price forward contracts. This implies that the 300 MW supplier has sold 240 MWh, the 200 MW supplier has sold 160 MWh and the remaining 100 MW suppliers have sold 80 MWh. At the 650 MWh level of demand no supplier is pivotal relative to its forward market position, because the largest supplier has forward commitment of 240 MWh, yet the minimum amount of energy it must produce to serve system demand is 150 MWh. Consequently, it has no incentive to withhold output to drive the spot price up if in doing so it produces less than 240 MWh. If it produces less than 240 MWh, then it must purchase the difference between 240 MWh and its output from the short-term energy market at the prevailing price to meet its forward contract obligation.

At this level of forward contracting, the largest supplier only becomes pivotal relative to its forward contract obligations if the level of demand exceeds 740 MWh, which is considerably larger than 500 MWh, the level of demand that causes it to be pivotal in a short-term market with no fixed-price forward contracts, and only slight smaller than 800 MWh, maximum possible energy that could be produced with 800 MW of generation capacity. The higher the level of fixed-price forward contract coverage, the higher the level demand at which one or more suppliers becomes pivotal relative to its forward contract position.

Focusing on the development of a long-term forward market has an additional dynamic benefits to the performance of short-term energy markets. If all suppliers have significant fixed-price forward contract commitments then all suppliers share a common interest in minimizing the cost of supplying these forward contract commitments, because each supplier always has the option purchase energy from the short-term market as opposed to supply this energy from its generation units. The dynamic benefit comes from the fact that at high levels of forward contracting the operating efficiency gains from re-structuring described above will be translated into spot prices.

Although the initial forward contracts signed between retailers and suppliers did not incorporate these expected efficiency gains in the prices charged to retailers, subsequent rounds of fixed-price forward contracts signed will incorporate the knowledge that these efficiency gains were achieved.

It is very important to emphasize that the initial round of forward contracting cannot capture these dynamic efficiency gains in the prices that retailers must pay, because these efficiency gains will not occur unless significant fixed-price forward contracting takes place. Moreover, this required amount of fixed-price forward contracting will not take place unless suppliers receive sufficiently high fixed-price forward contract prices to compensate them for giving up the short-term market revenues they could expect to receive if they did not sign the forward contracts. This difference between expected future short-term prices with and without high levels of fixed-price contracting can be very large.

An illustration of this point comes from the California market during the winter of 2001. Forward prices for summer 2001 deliveries were approximately $300/MWh. Those for summer 2002 deliveries were approximately $150/MWh and those summer 2003 and were approximately $45/MWh. Prices in summer 2001 were that high because by signing a fixed-price forward contract to supply during that time mean giving up significant opportunities to earn high prices in the short-term energy market. Forward prices for summer 2002 were half as high as those for summer 2001 because all supplies recognized that more new capacity and potentially more existing hydroelectric capacity could compete to supply energy to the short-term energy market in summer 2002 than in summer 2001. By the winter 2001, hydro conditions for summer 2001 have largely been determined, whereas those for summer 2002 are still very uncertain. Finally, the prices for summer 2003 were significantly lower, because suppliers recognized that a substantial amount of new generation capacity could come on line to compete in the short-term energy market by the summer of 2003. For this reason, suppliers expected that there would be few opportunities to exercise substantial unilateral market power in the short-term energy market during the summer of 2003, so they did not have to be compensated with a high energy price to sign a fixed-price forward contract to provide energy during the summer of 2003.

The second half of this story is that after the State of California signed significant fixed-price long-term forward contracts with suppliers at prices that reflected these forward market prices, short-

term market prices during the Summer of 2001 reflected low levels of unilateral market power despite the fact that hydroelectric energy conditions in the WECC were not appreciably different from those during the Summer of 2000. A major cause of these short-term market outcomes is the high level of fixed-price forward contract commitments many suppliers had to supply energy to California LSEs.

The above discussion provides strong evidence against the argument that getting the short-term market design right is the key to workably competitive short-term energy markets. Without significant coverage of final demand with fixed-price forward contracts it is virtually impossible to limit the opportunities for suppliers to exercise substantial unilateral market power in the short-term energy market during intermediate to high demand periods. In addition, those who argue that retailers should delay signing long-term forward contracts until the spot market become workably competitive are likely to be waiting an extremely long time. This discussion also demonstrates why, at least for the initial rounds of forward contracting between retailers and suppliers, it is extremely difficult to capture the operating efficiencies gains from restructuring in the forward contract prices. This is another reason for beginning any restructuring process with the vesting contracts that immediately set motion the incentive to translate operating efficiency gains into short-term wholesale prices.

7.2. Inadequate Amounts of Divestiture

A number of re-structuring processes have been plagued by inadequate amounts of divestiture or an inadequate process for divesting generation units from the incumbent vertically-integrated monopoly. Typically, political constraints make it extremely difficult to separate the former state-owned companies into a sufficiently large number of suppliers. This leads to a period when existing suppliers are able to exercise substantial unilateral market power in the short-term energy market, which then leads to calls for regulatory intervention. If the period of time when these suppliers are able exercise unilateral market power is sufficiently long, the regulator either successfully implements further divestiture or some other form of regulatory intervention takes place.

The England and Wales restructuring process followed this pattern. Initially, the fossil fuel capacity of the original state-owned vertically integrated utility, National Power, was sold off to two privately owned companies, the newly privatized National Power and PowerGen, with the nuclear

capacity of original National Power initially retained in a government owned-company. This effectively created a tight duopoly market structure in the England and Wales market, which allowed substantial unilateral market power to be exercised, once a significant fraction of the initial round of vesting contracts expired. Eventually the regulator was able to implement further divestitures of generation capacity from the two fossil fuel suppliers, and the high short-term prices that reflected significant unilateral market power triggered new entry by combined-cycle gas turbine (CCGT) capacity. At the same time calls for reform of the original England and Wales market design were justified based on the market power exercised by the two large fossil fuels suppliers. A strong case can be made that both the substantial amount of unilateral market power exercised from mid-1993 onwards and the subsequence expense of implementing the New Electricity Trading Arrangements (NETA) could have been avoided had more divestiture taken place at the start of the wholesale market.

New Zealand is an extreme example of insufficient divestiture at the start of the wholesale market regime. The Electricity Company of New Zealand (ECNZ) the original state-owned monopoly owned more than 95% of the generation capacity in New Zealand. Contact Energy, another state-owned entity was given 30% of this generation capacity at the start of the wholesale market. However, this duopoly market structure was thought to have market power problems and the amount of generation capacity owned by the largest stated-owned firm, virtually all of which was hydroelectric capacity, was thought to discourage needed private generation investment. Consequently, further divestiture of generation capacity from ECNZ was then implemented.

The poor experience of California with the divestiture process was not the result of an inadequate amount of divestiture, but how it was accomplished. First and foremost, the divested assets were sold without vesting contracts which would have allowed the three investor-owned utilities to buy a substantial fraction of the expected output of these units are a price set by the California Public Utilities Commission. As discussed in Wolak (2003b) the lack of substantial fixed-price forward contracts between these new suppliers and the three major California retailers created substantial opportunities for the owners of the divested assets to exercise substantial unilateral market power in California's short-term energy markets starting in June 2000 because the availability of hydroelectric energy was significantly less than the levels in 1998 and 1999. A second problem with the divestiture of generation assets in California is that these units where

typically purchased in tight geographic bundles, which significantly increased the local market power problem faced by California.

There appears to be one divestiture success story–the Victoria Electricity Supply Industry in Australia. The Victorian government decided to sell off all generation assets on a plant-by-plant basis.[8] Despite a peak demand in Victoria of approximately 7,500 MW and only three sizeable suppliers, each of which own one large coal-fired generation plant, the short-term energy market has been remarkably competitive since it began in 1994. Wolak (1999) describes the performance of the Victoria market during its first four years of operation.

Inadequate amounts of divestiture can also make achieving an economically reliable transmission network in the sense of Section 5.4 significantly more expensive. Comparing two otherwise identical wholesale markets, except that one has substantial amounts of transmission capacity interconnecting all generation units and load centers and the other has the minimum amount of transmission capacity interconnecting generation units and load centers, the former market is likely to be able to achieve acceptable levels of wholesale market performance with less divestiture. The market with a substantial amount of transmission capacity will allow more generation units to compete supply electricity at every location in the transmission network. This logic implies the following two conclusions. First, the amount of divestiture necessary to achieve a desired level of competitiveness of the wholesale market outcomes depends on the characteristics of the transmission network. Second, the economic reliability of a transmission network in the language of Section 5.4 depends on the concentration and location of generation ownership. More concentration of generation ownership implies that a more extensive and higher-capacity transmission network is necessary to achieve the same level of competitiveness of wholesale market outcomes as would be the case with a less concentration of generation ownership. In this sense, less divestiture of generation capacity implies larger transmission network costs to attain the same level of competitiveness of wholesale market outcomes.

7.3. Lack of Effective Local Market Power Mitigation Mechanism

Although the need for an effective local market power mitigation mechanism has been discussed in detail, the crucial role this mechanism plays in limiting the ability of suppliers to

[8]Recall that generation plants are typically composed of multiple generation units at the same location.

exercise both systemwide and local market power has not been emphasized. Once again, the experience of California is instructive about the harm that can occur as a result of a poorly-designed local market power mitigation mechanism. On the other hand, the PJM wholesale electricity market is an excellent example of how short-term market performance can be enhanced by the existence of an effective local market power mitigation mechanism.

At the start of the California market there was no explicit local market power mitigation mechanism for units not governed by what were called Reliability Must-Run (RMR) contracts. These contracts were assigned to specific generation units thought to be needed to maintain system reliability even though short-term energy prices during the hours they were needed to run were insufficient to cover their variable costs plus a return to capital invested in the unit. All generation units without RMR contracts (non-RMR units) taken out-of-merit order because they were needed to meet solve a local reliability need, where eligible to be paid as-bid to provide this service, subject only to the bid cap on the energy market.[9]

As discussed earlier, system conditions can and do arise when virtually any generation unit owner, including a number of non-RMR unit owners, possess substantial local market power, or in engineering terms, they are the only unit able to meet a local reliability energy need. Once several non-RMR unit owners learned to predict when their unit was needed to meet a local reliability need, they very quickly began to bid at or near the bid cap on the ISO's real-time market to provide this service. This method for exercising local market power became so widespread that one market participant that owned several units at the same location, two of which where RMR units, is alleged to have delayed repairs on its RMR units in order to have the remaining non-RMR units be paid as-bid to provide the necessary local reliability energy. This was brought to the attention of FERC which required the unit owner to repay the approximately $8 million in additional profits earned from this strategy, but it imposed no further penalties. For more on this case, see FERC (2001).

This exercise of substantial local market power enabled by the lack of an effective local market power mitigation mechanism in California became extremely costly. Several commentators have argued that it inappropriately led FERC to conclude that California's zonal market design was

[9]A generation unit is said to be taken out of merit order if there are other lower cost units (or lower bid units) that can supply the necessary energy, but they are unable to do so because transmission constraints prevent their energy from reaching final demand.

fatally flawed, despite the fact that zonal-pricing market designs are the dominant congestion management mechanism around the world. A case could be made that if California had a local market power mitigation mechanism similar to that in PJM or in several other zonal-pricing markets around the world, there would have been very few opportunities for suppliers to exercise the amount local market power that led FERC to its conclusion.

The PJM local market power mitigation mechanism is an example of an effective local market power mitigation mechanism. It applies to all units located in the PJM control area on a prospective basis. If the PJM ISO determines that a unit possesses substantial local market power during an hour, then that unit's bid is typically mitigated to a regulated variable cost in the day-ahead and real-time price-setting process. There are two other options available that can selected for the mitigated bid level, but this regulated variable cost is the most common choice by generation unit owners. Under the PJM mechanism, a supplier is deemed to possess local market power worthy of bid mitigation if additional energy is needed from this generation unit to resolve a transmission constraint within one of the small number of pre-designed geographic regions of the PJM control area. Wolak (2002) describes the generic the local market power problem in more detail and describes the details of the PJM local market power mitigation mechanism.

It is not difficult to imagine how the California market would have functioned if it had the PJM local market power mitigation mechanism from the start of the market. All suppliers taken to resolve local reliability problems would be paid a regulated variable cost, instead of as-bid up to the bid cap on the spot market, for this additional energy. The costs to resolve local reliability constraints would have been substantially lower and very likely not to have risen to a high enough level to cause alarm at FERC. This comparison of the PJM versus California experience with local market power mitigation mechanisms serves as a cautionary tale to market designers who fail to adequately address the local market power mitigation problem.

7.4. Lack of a Credible Bid or Price Cap on the Wholesale Market

Virtually all bid-based wholesale electricity markets have explicit or implicit bid caps. The proper level of the bid cap on the wholesale electricity market is largely a political decision, as long as it is set above the variable cost of the highest cost unit necessary to meet the annual peak demand. However, there is an important a caveat associated with this statement that is often not appreciated. In order for a bid cap to be credible, the ISO must have a pre-specified plan that it will implement

if there is insufficient generation capacity bid into the real-time market at or below the bid cap to meet real-time demand. Without this there is an extreme temptation for suppliers that are pivotal or nearly pivotal relative to their forward market positions in the short-term energy market to test the credibility of bid or price cap, and this can lead to an unraveling of the formal market mechanism.

There is an inverse relationship between the level of the price cap on the spot market that can be credibly maintained and the necessary amount of final demand that must be covered by fixed-price forward contracts for energy. Lower levels of the bid cap on the spot market for energy require higher levels of coverage of final demand with fixed-price forward contracts in order to maintain the integrity of the bid cap on the energy or ancillary services market. For example, the experience of the California market since the winter of 2002 has shown that a bid cap of $250/MWh does not impose significant reliability problems or degrade the efficiency of the spot market if virtually all of the demand is covered by fixed-price forward hedging arrangements.

If the bid cap is set too low for the level of forward contracts, then it is possible for system conditions to arise when one or more suppliers have an incentive to test the integrity of the bid cap on the spot market, by bidding in excess of the price cap. The ISO operators are faced with the choice blacking out certain customers in order to maintain the integrity of the transmission network, or paying suppliers their bids to provide the necessary energy. If the operators make the obvious choice of paying these suppliers their bids, other market participants will quickly find this out, which encourages them to raise their bids above the cap and the formal wholesale market begins to unravel.

System conditions when suppliers had the opportunity to test the integrity of the bid cap arose frequently during the period June 2000 to June 2001 because only a very small fraction of final demand was covered in fixed price forward contracts. Maintaining the credibility of a relatively low bid cap of say twice to three times variable cost of the highest cost unit in the system, requires that the regulatory process mandate fixed-price forward contract coverage of final demand at a very substantial fraction, certainly more than 90%, of final demand.

It is important to emphasize that this level of forward contracting must be mandated if a low bid cap is to be credible. Without this requirement, retailers have an incentive to rely on the short-term market and the protection against high short-term prices provided by the relatively low bid cap for their wholesale energy purchases, rather than voluntarily purchase sufficient fixed-priced

forward contracts to maintain the credibility of the bid cap.

7.5. Inadequate Retail Market Infrastructure

This section describes inadequacies in the physical and regulatory retail market infrastructure in many wholesale markets that can limit the competitiveness of the wholesale market. The first is the lack of interval metering necessary for final consumers to be active participants in the wholesale market. The second is the asymmetric treatment of load and generation by the state regulatory process. The lack of interval meters and asymmetric treatment of load and generation creates circumstances where final demand has little ability or incentive to take actions to enhance the competitiveness of wholesale market outcomes.

Virtually all existing meters for small commercial and residential customers in the US only capture total electricity consumption between consecutive meter readings. In the US, meters for residential and small business customers are usually read on a monthly basis. This means that the only information available to an electricity retailer about these customers is their total monthly consumption of electricity. In order to determine the total monthly wholesale energy and ancillary services cost to serve this customer, this monthly consumption is usually distributed across hours of the month according to a representative load shape proposed by the retailer and approved by the state regulator. For example, let q(i,d), denote the consumption of the representative consumer in hour i of day d. A customer with monthly consumption equal to Q(tot) is assumed to have consumption in hour i of day equal to:

$$qp(i,d) = \frac{q(i,d)Q(tot)}{\sum_{d=1}^{D}\sum_{i=1}^{24} q(i,d)}$$

This consumer's monthly wholesale energy bill is computed as

$$Monthy\ \ Wholesale\ \ Energy\ \ Bill = \sum_{d=1}^{D}\sum_{i=1}^{24} qp(i,d)p(i,d),$$

where p(i,d) is the wholesale price in hour i of day d. This expression can be simplified to P(avg)Q(tot), by defining P(avg) as:

$$P(avg) = \frac{\sum_{d=1}^{D}\sum_{i=1}^{24} p(i,d)qp(i,d)}{\sum_{d=1}^{D}\sum_{i=1}^{24} q(i,d)}$$

Despite this attempt to allocate monthly consumption across the hours of the month, in the end the consumer faces the same wholesale energy price for each KWh consumed during the month. If a customer maintained the same total monthly consumption but shifted it from hours with very high wholesale prices to those with low wholesale prices, the customer's bill would be unchanged.

Without the ability to record a customer's consumption on an hourly basis it is impossible to implement a pricing scheme that allows the customer to realize the full benefits of shifting his consumption from high-priced hours to low-priced hours. In a wholesale market the divergence between P(avg) and the actual hourly wholesale price can be enormous. For example, during the year 2000 in California, P(avg) was equal to approximately 10 cents/KWh despite the fact that the price paid for electricity often exceeded 75 cents/KWh and was as high as $3.50/KWh for a few transactions. In contrast, under the vertically-integrated utility regime, the utility received the same price for supplying electricity that the final customer paid for every KWh sold to that customer.

The installation of hourly meters would allow a customer to pay prices that reflect hourly wholesale market conditions for its electricity consumption during each hour. A customer facing an hourly wholesale price of $3.50/KWh for any consumption in that hour in excess of his forward market purchases would have a very strong incentive to cut back during that hour. This incentive extends to reductions in consumption below this customer's forward market purchases, because any energy not consumed below this forward contract quantity is sold at the short-term market price of $3.50/KWh.

The importance of recording consumption on an hourly basis for all customers can be best understood by recognizing that a 1 MWh reduction in electricity consumption is equivalent to a 1 MWh increase in electricity production assuming that both the 1 MWh demand decrease and 1 MWh supply increase are provided with the same response time and at the same location in the transmission grid. Because these two products are identical, in a world with no regulatory barriers to active demand side participation, arbitrage should force the prices paid for both products to be equal.

Virtually all customers in the US with hourly meters still have the option to purchase all of their electricity at a retail price that does not vary with hourly system conditions. All customers without interval meters have this same option. The supply-side analogue to this option to purchase as much electricity as the customer wants at a fixed price is not available to generation unit owners.

The default price a generation unit owner faces is the real-time wholesale price. If the supplier would like to receive a different price for its output, then it must sign a hedging arrangement with another market participant. To provide incentives for final consumers to manage wholesale price risk, they must also pay a default wholesale price equal to the real-time wholesale price. No consumer needs to pay this real-time price. If the consumer would like to pay a different price then it must sign a hedging arrangement with another market participant. Wolak (2007b) presents a simple model that shows if final consumers have the option to purchase as much as they want at a fixed retail price, this can destroy their incentive to manage their real-time price risk through altering their consumption in response to short-term prices.

To justify the existence of this option for consumers to purchase all of their consumption at a fixed price, state regulators will make the argument that customers must be protected from volatile short-term wholesale prices. However, this logic falls prey to the following economic reality, over the course of the year the total amount of revenues recovered from retail consumers after transmission, distribution and retailing charges have been subtracted must be sufficient to pay total wholesale energy purchase costs over that year. If this constraint is violated the retailer will earn a loss or be forced into bankruptcy unless some other entity makes up the difference. Consequently, consumers are not shielded from paying volatile wholesale prices. They are simply prevented from reducing their annual electricity bill by reducing their consumption during the hours when wholesale prices are high and increasing their consumption when wholesale prices are low.

A number of observers complain that retail competition provides few benefits to final consumers and does little to increase the competitiveness of wholesale market outcomes. Joskow (2000b) provides an extremely persuasive argument for this position. If retail competition is introduced without hourly metering and with a fixed default retail price, then it becomes extremely difficult to refute his argument.

Without hourly metering and a default retail price that passes through the hourly wholesale price it is difficult to see how retail competition can benefit electricity consumers. The logic for this view follows. Competition among firms occurs because one firm believes that it can better serve the needs of consumers than firms currently in the industry. These firms succeed either by offering an existing product at a lower cost or by offering new product that serves a previously unmet consumer need. Consider the case of electricity retailing without hourly meters. The only

information each retailer has is the customer's monthly consumption of electricity and some demographic characteristics that might be useful for predicting its monthly load shape, the q(i,d) described above. The dominant methodology for introducing retail competition is load-profile billing to the retailer for the hourly wholesale energy purchases necessary to serve each customer's monthly demand. This scheme implies that all competitive retailers receive the same monthly wholesale energy payment (for the wholesale electricity it allows the incumbent retailer to avoid purchasing on this customer's behalf) for each customer of a given type that they serve. Customer types are distinguished by a representative load shape and monthly consumption level.

Under this mechanism, competitors attract customers from the incumbent retailer by offering an average price for energy each month, P(avg) as defined above, that is below the value offered by other retailers. The inability to measure this customer's consumption on an hourly basis implies that competition between electricity retailers takes place on a single dimension, the monthly average price they offer to the consumer. The opportunities for retailers to exploit competitive advantages relative to other retailers under this mechanism are severely limited. Moreover, this mechanism for retail competition also always requires asymmetric treatment of the incumbent retailer relative to other competitive retailers. Finally, the state PUC must also continue to have an active role in this process because it must approve the representative load shapes used to compute P(avg) for each customer class.

With hourly metering and a default price that passes through the hourly wholesale price, retail competition has the greatest opportunity to provide tangible economic benefits. Competition to attract customers can now take place along as many as 744 dimensions, the maximum number of hours possible in a one month. A retailer can offer a customer as many at 744 different prices for a monthly period. Producers can offer a enormous variety of nonlinear pricing plans that depend on functions of their consumption in these 744 hours. Retailers can now specialize in serving certain load shapes or offering certain pricing plans as their way to achieve a competitive advantage over other retailers.

Hourly meters allows retailers to use retail pricing plans to match their retail load obligations to their hourly pattern of electricity purchases. Rather than having to buy a pre-determined load shape in the wholesale market, retailers can instead buy a less expensive load shape and use their retail pricing plan to offer significantly lower prices in some hours and significantly higher prices

in other hours to cause their retail customers to match this load shape yet achieve a lower average monthly retail electricity bill. This is possible because the retailer is able to pass on the lower cost of its wholesale energy purchases in the average hourly retail prices it charges the consumer.

Wolak (2007b) suggests a process for transitioning to universal hourly metering of customers and allowing retail competition in a region once the necessary metering has been installed. This scheme has the additional advantage of eliminating the need for asymmetric treatment of the incumbent retailer versus competitive retailers. Because every consumer's consumption is available at the level of time aggregation that wholesale electricity is bought and sold, there is no need for the regulator to set representative load shapes for any customer.

8. Explaining the US Experience with Electricity Industry Restructuring

This section uses the results of the previous four sections to diagnose the underlying causes of the disappointing performance of re-structured wholesale markets relative to the former vertically-integrated utility regime in the US. This is experience is compared to that of a number of other industrialized countries to better understand whether improvements in market performance in the restructured regime are possible in the US, or if industry restructuring in the US is doomed to be an extremely expensive experiment.

8.1. Federal versus State Regulatory Conflict

Rather than coordinating wholesale and retail market policies to benefit wholesale market performance, almost the opposite has happened in the US. State PUCs have designed retail market policies that attempt to maintain regulatory authority over the electricity supply industries in their state as FERC's authority grows. Retail market policies consistent with fostering a competitive wholesale market may appear to state PUCs has giving up regulatory authority. For example, making the default rate all retail customers pay equal to the real-time price, appears to be giving up on the state PUC's ability to protect consumers from volatile wholesale prices. Introducing retail competition also appears to be giving up the state PUC's the authority to set retail prices.

The vertically-integrated, regulated-monopoly regime in the US limited opportunities for conflicts between state and federal regulators. As noted earlier, this regime involved few short-term interstate wholesale market transactions. State regulators also had a dominant role in the transmission and generation capacity-planning decisions of the investor-owned utilities they regulated.

As discussed earlier The Federal Power Act requires that FERC set "just and reasonable" wholesale electricity prices. The following passage from the Federal Power Act clarifies the wide ranging authority FERC has to fulfill its mandate.

> Whenever the Commission, after a hearing had up its own motion or upon complaint, shall find that any rate, charge, or classification, demand, observed, charged or collected by any public utility for transmission or sale subject to the jurisdiction of the Commission, or that any rule, regulation, practice, or contract affected such rate, charge, or classification is unjust, unreasonable, unduly discriminatory or preferential, the Commission shall determine the just and reasonable rate, charge, classification rule, rule, regulation, practice or contract to be thereafter observed and in force, and shall fix the same by order (Federal Power Act).

Historically, just and reasonable prices are those that recover all prudently incurred production costs, including a return on capital invested.

For more than sixty years FERC implemented its obligations to set just and reasonable rates under the Federal Power Act by regulating wholesale market prices. During the 1990s, based on the belief that if appropriate criteria were met, "market-based rates" could produce lower prices and a more efficient electric power system, FERC changed its policy. It began to allow suppliers to sell wholesale electricity at market-based rates but, consistent with FERC's continuing responsibilities under the Federal Power Act, only if the suppliers could demonstrate that the resulting prices would be just and reasonable. Generally, FERC allowed suppliers to sell at market-based rates if they met a set of specific criteria, including a demonstration that the relevant markets would be characterized by effective competition. FERC retains this responsibility when a state decides to introduce a competitive wholesale electricity market. In particular, once FERC has granted suppliers market-based pricing authority it has an ongoing statutory responsibility to ensure that these market prices are just and reasonable.

The history of federal oversight of wholesale electricity transactions described above demonstrates that FERC has a very different perspective on the role of competitive wholesale markets than state PUCs or state policymakers. This difference is due in large part to the pressures put on FERC by the entities that it regulates versus the pressures put on state PUCs and policymakers by the entities they regulate. The merchant power producing sector has been very supportive of FERC's goal of promoting wholesale markets. These companies have taken part in

a number of lawsuits and legislative efforts to expand the scope of federal jurisdiction over the electricity supply industry.

In contrast, state PUCs face a very different set of incentives and constraints. First, for more than 50 years, state PUCs have set the retail price of electricity and managed the process of determining the magnitude and fuel mix of new generation investments by the investor-owned utilities within their boundaries. This paternal relationship between the PUC and the firms that it regulates makes it extremely difficult to implement the physical and regulatory infrastructure necessary for a successful wholesale market.

Neither the state PUC nor the incumbent investor-owned utility benefits from the introduction wholesale competition. The state PUC loses the ability to set retail electricity prices and the investor-owned utility faces the prospect of losing customers to competitive retailers. It is difficult to imagine a state regulator or policymaker voluntarily giving up the authority to set retail prices which can benefit certain customer classes and harm other customer classes. Because every citizen of a state consumes some electricity, the price-setting process is an irresistibly tempting opportunity for regulators and state policymakers to pursue social goals in the name of industry regulation.

The introduction of wholesale competition also limits the scope for the PUC and state policymakers to determine the magnitude and fuel mix of new generating capacity investments. Different from the former regulated regime where the PUC and state government played a major role in determining both the magnitude of new capacity investments and the input fuel for this new investment, in the wholesale market regime, this decision is typically made by independent, non-utility power producers.

For these reasons, the expansion of wholesale competition and the creation of the retail infrastructure necessary to support it directly conflict with many of the goals of the state PUCs and incumbent investor-owned utilities. Because it is a former monopolist, the incumbent investor-owned utility only stands to lose retail customers as a result of the implementation of effective retail competition. It is usually among the largest employers in the state, so it is often able to exert influence over the state-level regulatory process to protect its financial interests. Because the state PUC loses much of its ability control the destiny of the electricity supply industry within its boundaries when wholesale and retail competition is introduced, the incumbent investor-owned

utility finds a very sympathetic ear to arguments against adopting the retail market infrastructure necessary to support a wholesale market that benefits final consumers.

FERC's statutory responsibility to take actions to set just and reasonable wholesale rates, provides state PUCs with the opportunity to appear to fulfill their statutory mandate to protect consumers from unjust prices, yet at the same time serve the interests of their incumbent investor-owned utilities. The state can appease the incumbent investor-owned utility's desire to delay or prohibit retail competition by relying on FERC to protect consumers from unjust and unreasonable wholesale prices though regulatory interventions such as price caps or bid caps on the wholesale market. However, the events of May 2000 to May 2001 in California have emphasized, markets do not always set just and reasonable rates, and FERC's conception of policies that protect consumers from unjust and unreasonable prices may be very different from those the state PUC and other state policymakers would like FERC to implement.

The lesson from California is that once a state introduces a wholesale market with a significant merchant generation segment–generation owners with no regulated retail load obligations–it gives up the ability to control retail prices. As discussed earlier California divested virtually all of its fossil-fuel generation capacity to five merchant suppliers with no vesting contracts. This is in sharp contrast to the experience of the eastern US wholesale markets in PJM, New England and New York, which were formed from tight power pools.[10] Typically the vertically-integrated utilities retained a substantial amount, if not all of their generation capacity in the wholesale market regime. Those that were required to sell generation capacity, did so with vesting contracts that allowed the selling utility to purchase energy from the new owner under long-duration fixed-price forward contracts. As a consequence of these decisions, the eastern ISOs had very few generation owners with substantial net long positions relative to their retail load obligations. Consequently, suppliers in these markets had less of an incentive and ability to exercise unilateral market power at all load levels, relative to California, where virtually all of the output of the non-utility generation sector was purchased in California's short-term energy and ancillary services markets.

---

[10] In the former vertically integrated regime, a power pool is a collection of vertically integrated utilities who decide to "pool" their generation resources to be dispatched by a single system operator to serve their joint demand.

8.2. Over Seventy Years of Regulating Privately-Owned Vertically Integrated Utilities

Another reason for the poor experience of the US relative to virtually all other countries in the world is the different starting points of the re-structuring process in the US versus other industrialized countries. Before restructuring in the US, there had been over 70 years of state-level regulatory oversight of privately-owned vertically integrated utilities. Recall the two tenets of state-level regulation described earlier are: (1) the obligation of the utility to serve all demand in its service territory at the regulated price, and (2) the requirement that the state PUC set a regulated price that allows the utility an opportunity to recover all prudently incurred costs to serve that demand. Once these regulated retail prices are set, a profit-maximizing utility wants to minimize the total costs of meeting this demand. This combination of effective state-level regulation and privately-owned profit-maximizing utilities has squeezed out much of the productive inefficiencies in the vertically-integrated utility's operations. Because the three eastern US markets started as tight power pools, it is also likely that this same mechanism operated to squeeze out many of the significant productive inefficiencies in the joint operation of the transmission network and generation units of the vertically-integrated utilities that were members of the power pool.

In contrast, wholesale markets in other industrialized countries such as England and Wales, Australia, New Zealand and the Nordic countries were formed from government-owned national or regional monopolies. As discussed earlier state-owned companies have significantly less incentive to minimize production costs than do privately-owned, profit-maximizing companies facing state-level regulatory oversight of their prices. These state-owned companies are often faced with political pressures to pursue other objectives besides least-cost supply of electricity to final consumers. They are often used to distribute political patronage in the form of construction projects or jobs within the company or to provide jobs in certain regions of the country. Consequently, the inefficiencies before re-structuring were likely to be far greater in the electricity supply industries in these countries or regions than in the US. Consequently, one explanation for the superior performance of the markets in these countries relative to the former vertically-integrated utility regime is that the potential benefits from restructuring were far greater in these countries, because there were more productive inefficiencies in the industries in these countries to begin with. In this sense, the relatively unimpressive performance of restructured markets in the US is the result of the combination of a relatively effective regulatory process and private ownership of the utilities. This

logic raises the important question of whether the major source of benefits in many of these industrialized countries is due to privatization of former state-owned utilities or the formation of a wholesale electricity market.

8.3. Increasing Amount of Intervention in Short-Term Energy Markets

Partially in response to the aftermath of the California Electricity Crisis, many aspects of wholesale market in the US have evolved to become very inefficient forms of cost-of-service regulation. One such mechanism that has become increasing the popular with FERC is the Automatic Mitigation Procedure (AMP) which is designed to limit the ability of suppliers to exercise unilateral market power in short-term energy markets. Bid adders for mitigated generation units are another FERC-mandated source of market inefficiencies.

The AMP mechanism uses a two-step procedure to determine whether to mitigate a generation unit. First, all generation unit owners have a reference price, typically based on accepted bids during what are determined by FERC to be competitive market conditions. If a suppliers bids in excess of this reference price by some preset limit, for example $100/MWh or 100% of the reference level, then this supplier violates the conduct test. Second, if this supplier's bid moves the market price by some preset amount, for example $50/MWh, then this bid is said to violate the impact test. A supplier's bid will be mitigated to its reference level if it violates the conduct and impact test. All US ISOs except PJM have a AMP mechanism in place.

Because the reference prices in the AMP mechanism are set based on the average of past accepted bids, there is a strong incentive for what is called "reference price creep" to occur. Accepted low bids can reduce a unit's reference price, which then limits the ability of the owner to bid high during system conditions when it able to move the market price through its unilateral actions. Consequently, this cost to bidding low during competitive conditions implies that the AMP mechanism may introduce more market inefficiencies than it eliminates, particularly in a market with a relatively low bid cap on the short-term energy market. Off-peak prices are higher than they would be in the absence of the AMP mechanism and average on-peak prices are not reduced sufficiently by the AMP mechanism to overcome these higher average prices during the off-peak hours.

The use of bid adders that enter into the day-ahead and real-time price-setting process have become increasing favored by FERC as a way to ensure that generation units mitigated by an AMP

mechanism or local market power mitigation mechanism earn sufficient revenues to remain financially viable. Before discussing the impact of these bid adders, it is useful to consider the goal of market power mitigation mechanisms: To produce locational prices that accurately reflect the incremental cost of withdrawing power at all locations in the network. Prices that satisfy this condition are produced by effective competition. An efficient price should reflect the incremental cost to the system of additional consumption at that location in the transmission network. A price that is above the short-term incremental cost of supplying electricity is inefficient because it can deter consumption whose value is greater than the cost of production, but below the price. Settting price equal to the marginal willingness of demand to curtail is economically efficient only if pricing at the variable cost of the highest cost unit operating would create an excess demand for electricity. When a generation unit owner bids above the unit's incremental cost, other more expensive units may be chosen to supply in the unit's place. Price-taking, profit-maximizing firms will choose to produce as long as the market price is above their incremental costs.

The goal of local market power mitigation is to induce an offer price from a generation unit with local market power equal to the one that would obtain if that unit faced sufficient competition. A unit that faces substantial competition would offer a price equal to its variable cost of supplying additional energy. When the LMPM mechanism is triggered, the offer price of such a unit is set to a regulated level. By the above logic, this regulated level should be equal to the ISO's best estimate of the unit's variable cost of supplying energy.

Although bid mitigation controls the extent to which offer prices deviate from incremental costs, bid adders, adding a $/MWh amount to the ISO's best estimate of the unit's minimum variable cost of operating, biases the offer price upwards to guarantee that mitigated offer prices will be noticeably higher than those from units facing substantial competition. Typically these bid adders are set at 10% of the unit's estimated variable cost. For units that are frequently mitigated, in terms of the fraction of their run hours, these bid adders can be extremely large, on the order to $40/MWh to $60/MWh in some ISOs, which is more than double the average price in many markets.

The use of a bid adder that is known to be larger than the generation unit's minimum variable cost contradicts the primary goal of the market design process. Generation units that face sufficient competition will bid close to their minimum variable cost. Combining these bids with mitigated bids

set significantly above their minimum variable cost of supplying energy will result in units facing significant competition being overused. One might think that a 10 percent adder is relatively small, but it is important to emphasize that if a 100 MW generation unit is operating 2000 hours per year with a 10 percent adder on top of a variable cost estimate of $50/MWh, this implies annual payments in excess of these variable costs of $1 million to that generation unit owner. In addition, this mitigated bid level will set higher prices for units located near this generation unit, further increasing the costs to consumers.

Frequently mitigated generation units are providing a regulated service, and for that reason they should be guaranteed recovery of all prudently incurred costs. But cost recovery need not distort market prices in periods or at locations where there is no other justification for them to rise above incremental costs. Consider a mitigated unit with a $60/MWh incremental cost and a $40/MWh adder that is applied in an hour of ample supply. The market will be telling suppliers with costs less than $100/MWh that they are needed and telling demand with a value of electricity less than $100/MWh to shut down. Neither outcome is desirable. FERC has articulated the belief that it is appropriate that some portion of the fixed costs of mitigated units be allowed to set market prices. In other words, such units should not just be allowed to recover their fixed costs for themselves, but those costs should be reflected in the prices earned by other non-mitigated units.

FERC is essentially arguing that prices should be set at long-run average cost, as they would in the long run in a competitive market. There are two problems with this view. The first is that the FERC would set prices to recover at least these average costs during all hours the unit operates. In a competitive market the high prices during certain periods would offset prices at incremental costs during the majority of hours with abundant supply. The average of all these resulting prices would trend toward long-run average cost. The adder approach sets the economically inefficient price all of the time, resulting the problems described above.

8.4. Transmission Network Ill-Suited for Wholesale Market

The legacy of state-ownership in other industrialized countries versus private-ownership with effective state-level regulation in the US implies that these industrialized countries began the re-structuring process with significantly more transmission capacity than did the US investor-owned utilities. In addition, the transmission assets of the former government monopoly were usually sold off as a single transmission network owner for the entire country, rather than maintained as separate

but interconnected transmission networks owned by the former utilities, as is the case in the US wholesale markets. Both of these factors argue in favor of the view that initial conditions in the transmission network in these industrialized countries was significant more likely to have an economically reliable transmission network for the wholesale market regime than that transmission networks in the US.

8.5. Too Many Carrots, Too Few Sticks

There are two ways to make firms do what the regulator wants them to do: (1) pay them money for doing it, or (2) pay them less money for not doing it. Much of the regulatory oversight at FERC has used the former solution, which implies that consumers are less likely to see benefits from a wholesale market.

A potential consumer benefit from a wholesale market is that all investments, no matter how prudent they initially seem, are not guaranteed full cost recovery. Generation units investments that turn out not to be needed to meet demand, do not receive full cost recovery. As with other markets, investors in these assets should bear the full cost of their "mistake," particularly if they also expect to receive all of the benefits associated with constructing new capacity when it is actually needed to meet demand. This investment "mistake" should be confined to the investor that decided to build the plant, not shared with all electricity consumers. Even if the entity that constructed the generation unit goes bankrupt, the generating facility is very unlikely to exit the market. Instead the new owners will be able to purchase the facility at less than the initial construction cost, reflecting the fact that the new generation capacity is not needed at that time. The unit will still be available to supply electricity consumers, the original owner just won't be the entity earning those revenues. The new owner is likely to continue to operate the unit, but with a significantly lower revenue requirement than the original investor. By allowing investors who invest in new generation capacity at what turns out to be the wrong time bear the cost of these decisions, consumers will have a greater likelihood of benefitting from wholesale competition.

A second way that FERC implicitly ends up paying suppliers more money to do what it wants, is the result of FERC's reliance on voluntary settlements among market participants. Because, as mentioned earlier, wholesale price regulation at FERC largely entailed approving terms and conditions negotiated under state-level regulatory oversight, FERC appears have drawn the mistaken impression that voluntary negotiation can be used to set regulated terms and conditions.

One way to characterize effective regulation is making firms do things they are able to do, but don't want to do. For example, the firm may be able to cover its production costs at a lower output price, but it has little interest in doing so if this requires greater effort from its management. Asking parties to determine the appropriate price that suppliers can charge retailers for wholesale power through a consensus among the parties present is bound to result in the party that is excluded from this process–final consumers–paying more. In order for consumers to have a chance of benefitting from wholesale competition, FERC must recognize this basic tenet of consensus solutions, and protect consumers from unjust and unreasonable prices.

10. Conclusions

It may be practically impossible to achieve the regulatory process in the US necessary for restructuring to benefit final consumers relative to the former vertically-integrated, regulated-monopoly regime. Much more so that in this regime, wholesale and retail market policies must be extremely well-matched in the restructured regime. Even in countries with the same entity regulating the wholesale and retail sides of the electricity supply industry, this is an extremely challenging task. For the US, with the historically adversarial relationship between FERC and state PUCs, presents an almost impossible challenge that has only been made more challenging by how FERC is generally perceived by state policymakers to have handled the California electricity crisis. These relationships appear to have improved in recent years as a result of number of positive changes at FERC, but there appears to be little common ground between FERC and many state PUCs concerning the best way forward with electricity industry restructuring. Virtually all states that could put on hold their restructuring plans have done so. A number are even attempting to push the clock bank in terms of the amount of wholesale competition relative to what currently exists within their boundaries.

The most prudent path forward for FERC appears to be to focus on enhancing the efficiency of the existing wholesale markets in the northeastern US, the midwest and California, rather than attempt to increase the number of wholesale markets. As should be clear from the previous section, a significant amount of outstanding market design issues remain, and a number of them do not have clear cut solutions. Dealing with these issues in the context of existing market designs appears to be the most prudent way forward.

References

(All references marked with (*) are available at http://www.stanford.edu/~wolak)

Averch, H. and Johnson, L. (1962), "Behavior of the Firm Under Regulatory Constraint," American Economic Review, December.

Awad, Mohamed; Casey, Keith E.; Geevarghese, Anna S.; Miller, Jeffrey C.; Rahimi, A. Farrokh; Sheffrin, Anjali Y.; Zhang, Mingxia; Toolson, Eric; Drayton, Glenn, Hobbs, Benjamin F.; and Wolak, Frank A. (2007) "Economic Assessment of Transmission Upgrades: Application of the California ISO Approach, available from http://www.stanford.edu/~wolak.

Borenstein, Severin, Bushnell, James, and Stoft, Steven (2000) "The Competitive Effects of Transmission Capacity in a Deregulated Electricity Industry," RAND Journal of Economics, Volume 31, Number 2, 294-325.

Borenstein, Severin, Bushnell, James, and Wolak, Frank A. (2002) "Measuring Market Inefficiencies in California's Restructured Wholesale Electricity Market," American Economic Review, December, 1367-1405.(*)

Borenstein, Severin (2005) "Wealth Transfers from Implementing Real-Time Retail Electricity Pricing" August, Center for the Study of Energy Markets Working Paper Number CSEMWP-147, available from http://www.ucei.berkeley.edu/pubs-csemwp.html

Bushnell, James and Saravia, Celeste (2002) "An Empirical Assessment of the Competitiveness of the New England Electricity Market," May, Center for the Study of Energy Markets Working Paper Number CSEMWP-101, available from http://www.ucei.berkeley.edu/pubs-csemwp.html

Bushnell, James and Wolfram, Catherine (2005) "Ownership Change, Incentives and Plant Efficiency: The Divestiture of U.S. Electric Generation Plants, March, Center for the Study of Energy Markets Working Paper Number CSEMWP-140, available from http://www.ucei.berkeley.edu/pubs-csemwp.html

Bushnell, James (2005) "Looking for Trouble: Competition Policy in the U.S. Electricity Industry" in Electricity Deregulation: Choices and Challenges (edited by) James Griffin and Steven Puller, University of Chicago Press: Chicago, IL.

Bushnell, James (2005) "Electricity Resource Adequacy: Matching Policies and Goals," August, Center for the Study of Energy Markets Working Paper Number CSEMWP-146, available

from http://www.ucei.berkeley.edu/pubs-csemwp.html

Charles River Associates (2004) "Statewide Pricing Pilot Summer 2003 Impact Analysis," Charles River Associates, 5335 College Avenue, Suite 26, Oakland, California 94618.

FERC (2001) "Order Approving and Stipulation and Consent Agreement, " (Issued April 30, 2001), AES Southland, Inc./Williams Energy Marketing & Trading Company, Docket No. IN01-3-001, United States of American, 95 FERC 61,167.

Fabrizio, Kira M., Rose, Nancy L., and Wolfram, Catherine (2007) "Do Markets Reduce Costs? Assessing the Impact of Regulatory Restructuring on U.S. Electric Generation Efficiency," American Economic Review, September, 1250-1278.

Hirst, Eric (2004) "US Transmission Capacity: Present Status and Future Prospects," Report prepared for Energy Delivery Group, Edison Electric Institute, and Office of Transmission and Distribution U.S. Department of Energy, June, available at http://electricity.doe.gov/documents/transmission_capacity.pdf.

Jarrell, Gregg A. (1978) "The Demand for State Regulation of The Electric Utility Industry," Journal of Law and Economics, pp. 269-295.

Joskow, "Inflation and Environmental Concern: Structural Change in the Process of Public Utilities Regulation," Journal of Law and Economics, October 1974.

Joskow, Paul (1987) "Productivity Growth and Technical Change in the Generation of Electricity," *The Energy Journal*, 8(1), 17-38.

Joskow, Paul (1989) "Regulatory Failure, Regulatory Reform, and Structural Change in the Electrical Power Industry," Brookings Papers on Economic Activity: Microeconomics, pp. 125-208.

Joskow, Paul (1997) "Restructuring, Competition and Regulatory Reform in the U.S. Electricity Sector, Journal of Economic Perspective, Summer, 11(3), 119-138.

Joskow, Paul (2000a) "Deregulation and Regulatory Reform in the U.S. Electric Sector," in Deregulation of Network Industries, Sam Peltzman and Clifford Winston (Editors), AEI-Brookings Joint Center for Regulatory Studies, pp. 113-188.

Joskow, Paul (2000b) "Why Do We Need Electricity Retailers? or Can You Get It Cheaper Wholesale?" Working Paper, available at http://econ-www.mit.edu/files/1127.

Joskow, Paul (2008) "Incentive Regulation in Theory and Practice: Electricity Distribution and

Transmission Networks," Chapter xx in this volume.

Joskow, Paul and Schmalensee, Richard (1983) *Markets for Power: An Analysis of Electric Utility Deregulation*. MIT Press: Cambridge, MA.

Joskow, Paul and Kahn, Edward (2002) "A Quantitative Analysis of Pricing Behavior In California's Wholesale Electricity Market During Summer 2000: The Final Word," The Energy Journal, December.

Laffont, Jean-Jacques, and Tirole, Jean (1991) "Privatization and Incentives," Journal of Law, Economics, and Organization, volume 7, Spring, 84-105.

Lee, Byung-Joo, (1995) "Separability Test for the Electricity Supply Industry," *Journal of Applied Econometrics*, 10, 49-60.

Mansur, Erin T. (2003) "Vertical Integration in Restructured Electricity Markets: Measuring Market Efficiency and Firm Conduct," October, Center for the Study of Energy Markets Working Paper Number CSEMWP-117, available from http://www.ucei.berkeley.edu/pubs-csemwp.html.

Megginson, William L. and Netter, Jeffry M. (2001) "From State to Market: A Survey of Empirical Studies of Privatization," Journal of Economic Literature, volume XXXIX, June, 321-389.

Patrick, Robert H. and Wolak, Frank A. (1997) "Estimating the Customer-Level Demand for Electricity Under Real-Time Market Prices," available from http://www.stanford.edu/~wolak.

Peltzman, Sam (1976) "Toward a More General Theory of Regulation," Journal of Law and Economics, number 2, 211-240.

Stigler, George (1971) "The Theory of Economic Regulation," Bell Journal of Economics and Management Science, volume 2, number 1, 3-22.

Shirley, Mary E. and Walsh, Patrick (2000) "Public versus Private Ownership: The Current State of the Debate," World Bank Policy Research Working Paper Number 2420, August.

Viscusi, W.Kip., Vernon, John M., and Harrington, Jr., Joseph E. (2001) Economics of Regulation and Antitrust (3rd Edition), The MIT Press, Cambridge, MA, 2001.

Wolak, Frank A. (1994), "An Econometric Analysis of the Asymmetric Information Regulator-Utility Interaction," Annales d'Economie et de Statistique, 34, 1994, 13-69.

Wolak, Frank A. and Patrick, Robert H. (1997) "The Impact of Market Rules and Market Structure

on the Price Determination Process in the England and Wales Electricity Market," February. (*)

Wolak, Frank A. (1999) "Market Design and Price Behavior in Restructured Electricity Markets: An International Comparison," in Competition Policy in the Asia Pacific Region, EASE Volume 8, Takatoshi Ito and Anne Krueger (editors) University of Chicago Press, 79-134.(*)

Wolak, Frank A. (2000a) "An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market," International Economic Journal, Summer, 1-40.(*)

Wolak, Frank A. (2000b) "Comments on the Office of Gas and Electricity Markets (Ofgem) License Condition Prohibiting Abuse of Substantial Market Power," submission to United Kingdom Competition Commission, July.(*)

Wolak, Frank A. (2001) "Designing a Competitive Electricity Market that Benefits Consumers, November (*).

Wolak, Frank A. (2002) "Competition-Enhancing Local Market Power Mitigation in Wholesale Electricity Markets" November. (*).

Wolak, Frank A. (2003a) "Measuring Unilateral Market Power in Wholesale Electricity Markets: The California Market 1998 to 2000," American Economic Review, May 2003, 425-430.(*)

Wolak, Frank A. (2003b) "Diagnosing the California Electricity Crisis," The Electricity Journal, August, 11-37. (*)

Wolak, Frank A. (2003c) "The Benefits of an Electron Superhighway," Stanford Institute for Economic Policy Research Policy Brief. November.(*)

Wolak, Frank A. (2003d) "Sorry, Mr. Falk: It's Too Late to Implement Your Recommendations Now: Regulating Wholesale Markets in the Aftermath of the California Crisis," The Electricity Journal, August, 50-55.

Wolak, Frank A. (2004) "Managing Unilateral Market Power in Wholesale Electricity," in *The Pros and Cons of Antitrust in Deregulated Markets*, edited by Mats Bergman, Swedish Competition Authority, xx-yy.

Wolak, Frank A. (2006) "Residential Customer Response to Real-Time Pricing: The Anaheim Critical-Peak Pricing Experiment," available from http://www.stanford.edu/~wolak.

Wolak, Frank A., (2007a) "Quantifying the Supply-Side Benefits from Forward Contracting in

Wholesale Electricity Markets," forthcoming, *Journal of Applied Econometrics*.

Wolak, Frank A., (2007b) "Managing Demand-Side Economic and Political Constraints on Electricity Industry Re-structuring Processes," forthcoming, *Stanford Law and Policy Review*.

Wolak, Frank A., Nordhaus, Robert and Shapiro, Carl, (2000) "Analysis of "Order Proposing Remedies for California Wholesale Electric Markets (Issued November 1, 2000)"," Market Surveillance Committee of the California Independent System Operator, December, at http://www.caiso.com/docs/2000/12/01/2000120116120227219.pdf

Wolfram, Catherine (1999) "Measuring Duopoly Power in the British Electricity Spot Market," American Economic Review, 89(4): 805-826.

Wolfram, Catherine (2005) "The Efficiency of Electricity Generation in the U.S. After Re-structuring," in Electricity Deregulation: Choices and Challenges (edited by) James Griffin and Steven Puller, University of Chicago Press: Chicago, IL.

Figure 1: Load Duration Curves for Victoria for 2000 to 2002

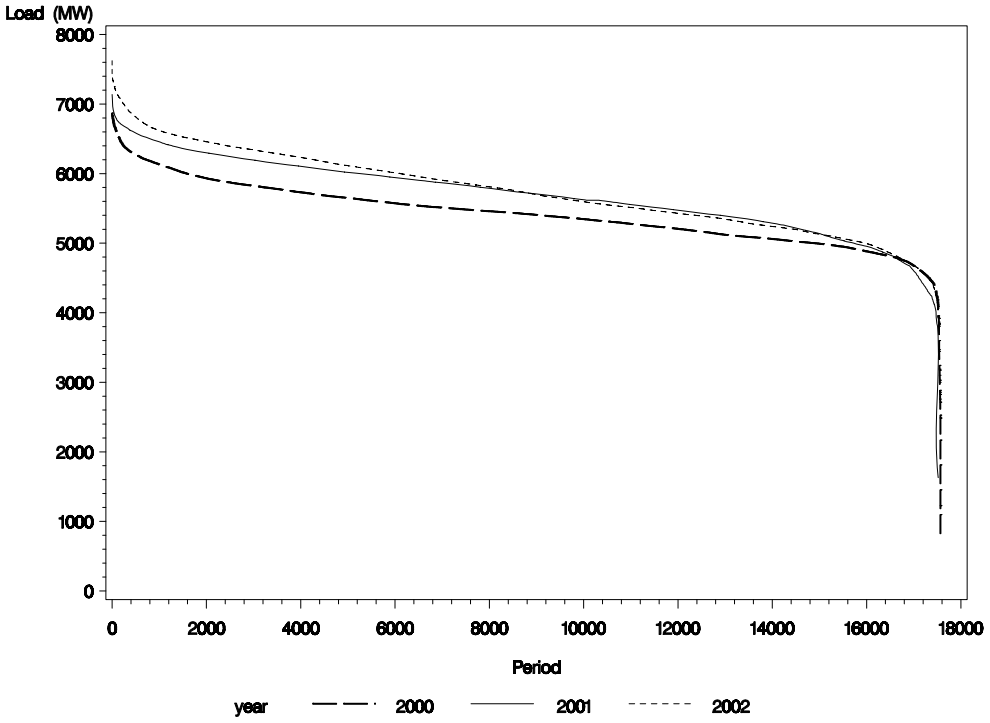

Figure 2: Annual Average Daily Pattern of System Load for Victoria for 2000 to 2002

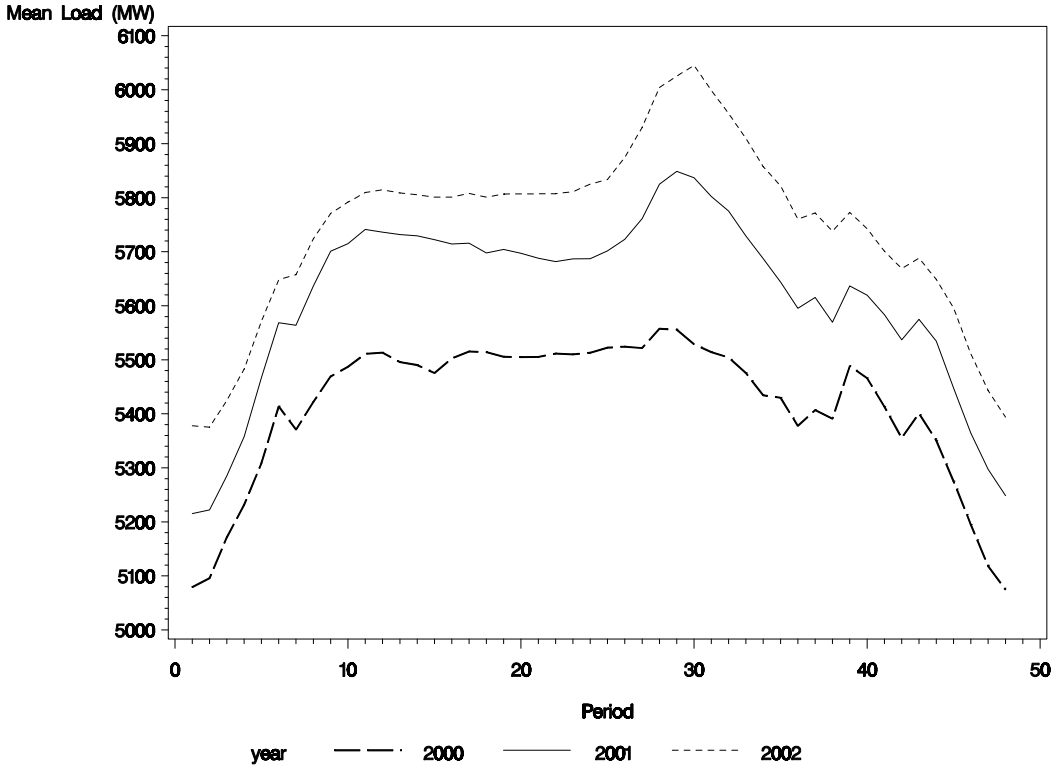

Figure 3: Annual Average Daily Pattern of Output for Yallourn Electricity Generation Plant

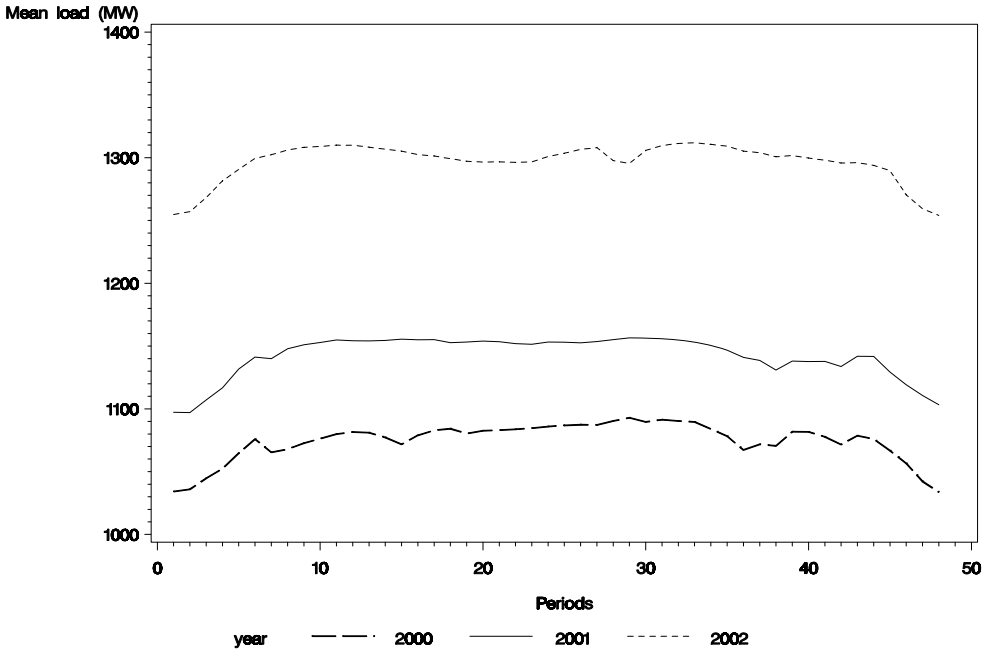

Figure 4: Annual Average Daily Pattern of Output for Valley Power Electricity Generation Plant

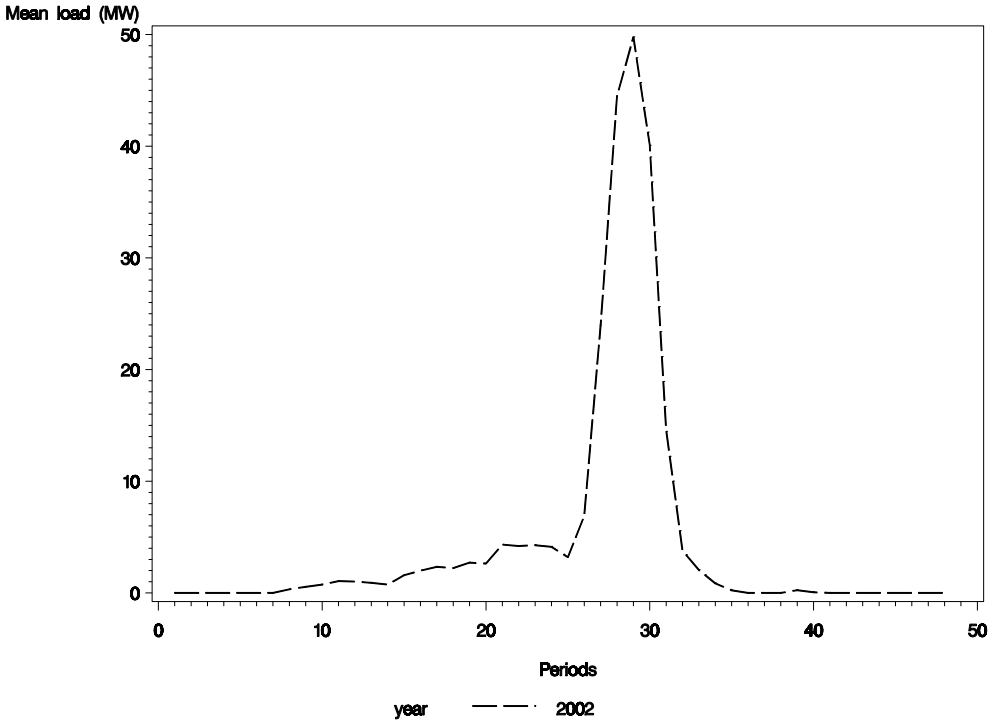

Figure 5: Load Duration Curve for Highest 500 Half-Hours for Victoria from 2000 to 2002

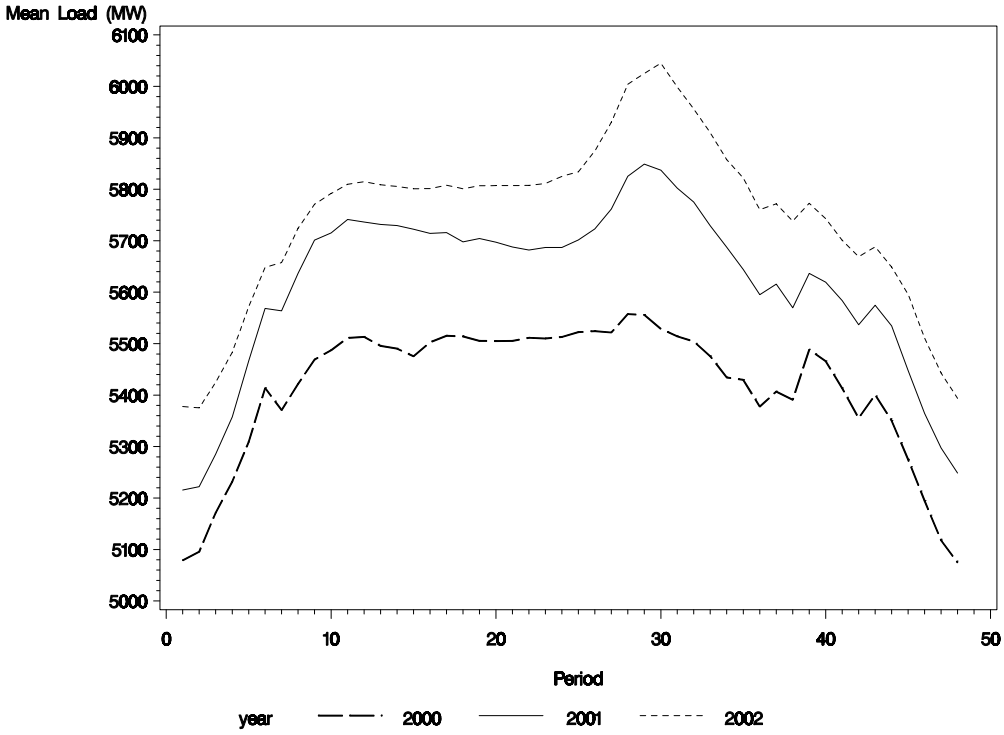

Figure 6: Annual Average Half-Hourly Prices for Victoria from 2000 to 2002

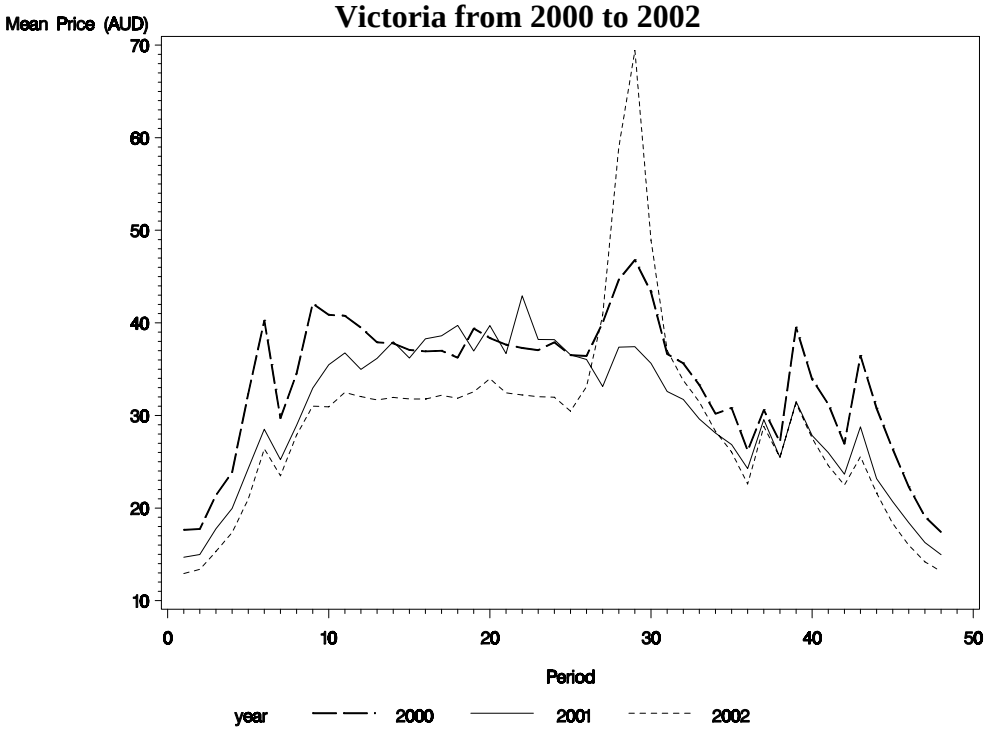

Figure 7: Power Flows in a Three Node Network

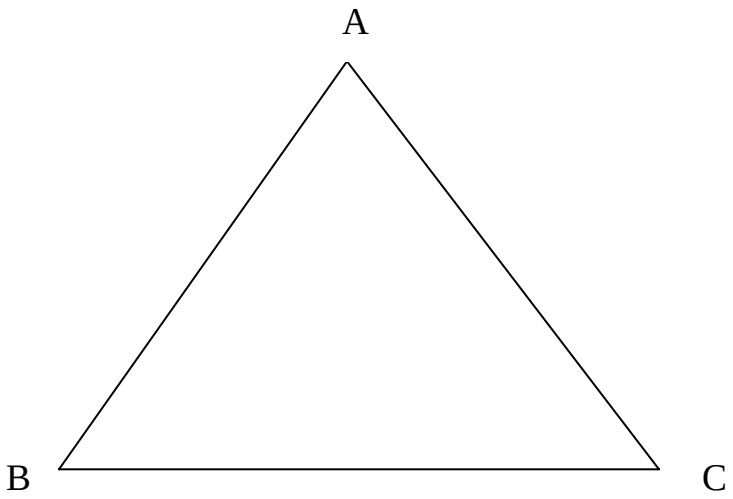

Figure 8: Residual Demand Elasticity and Profit-Maximizing Behavior

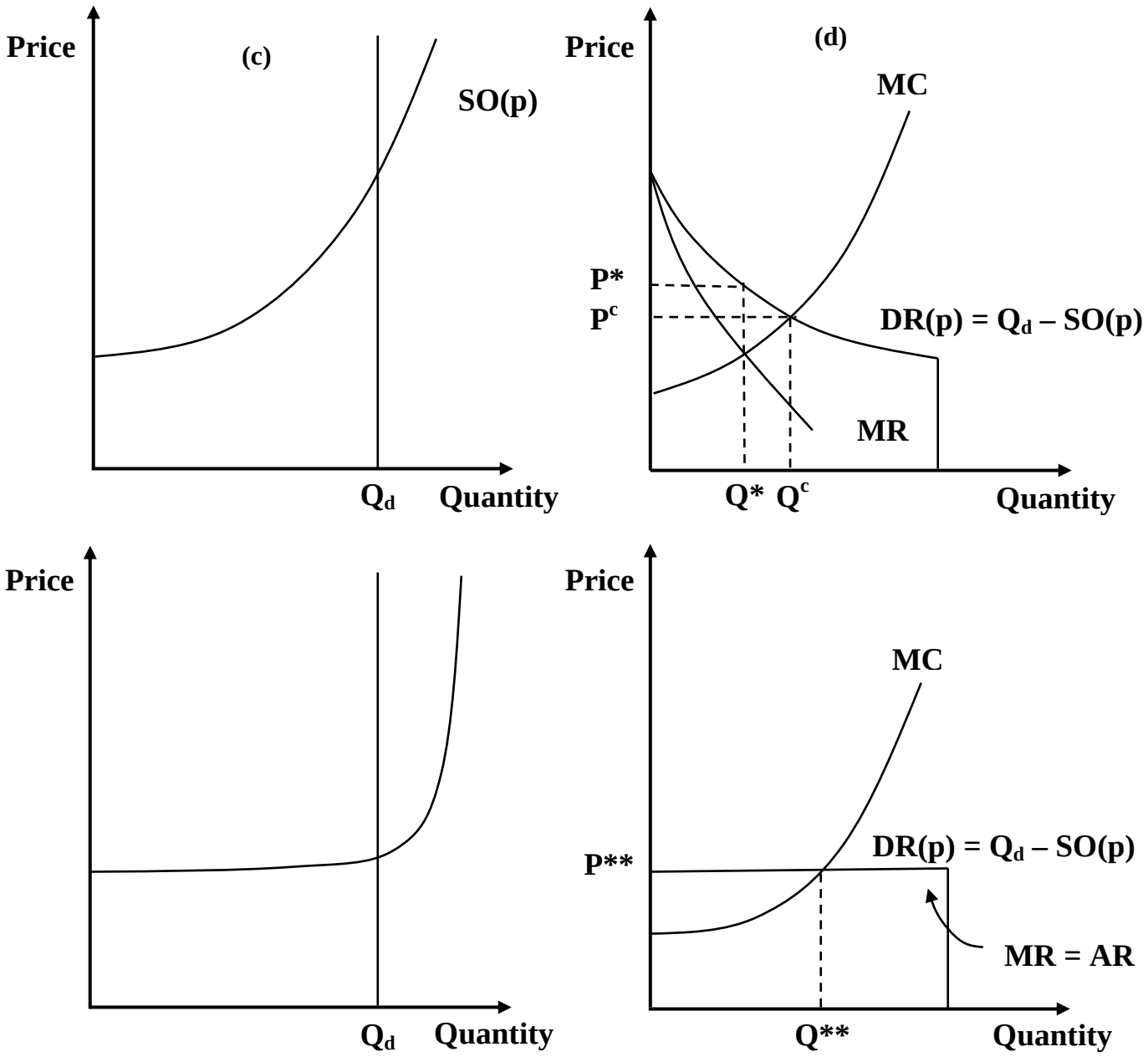

Figure 9: Welfare Loss from Inefficient Production

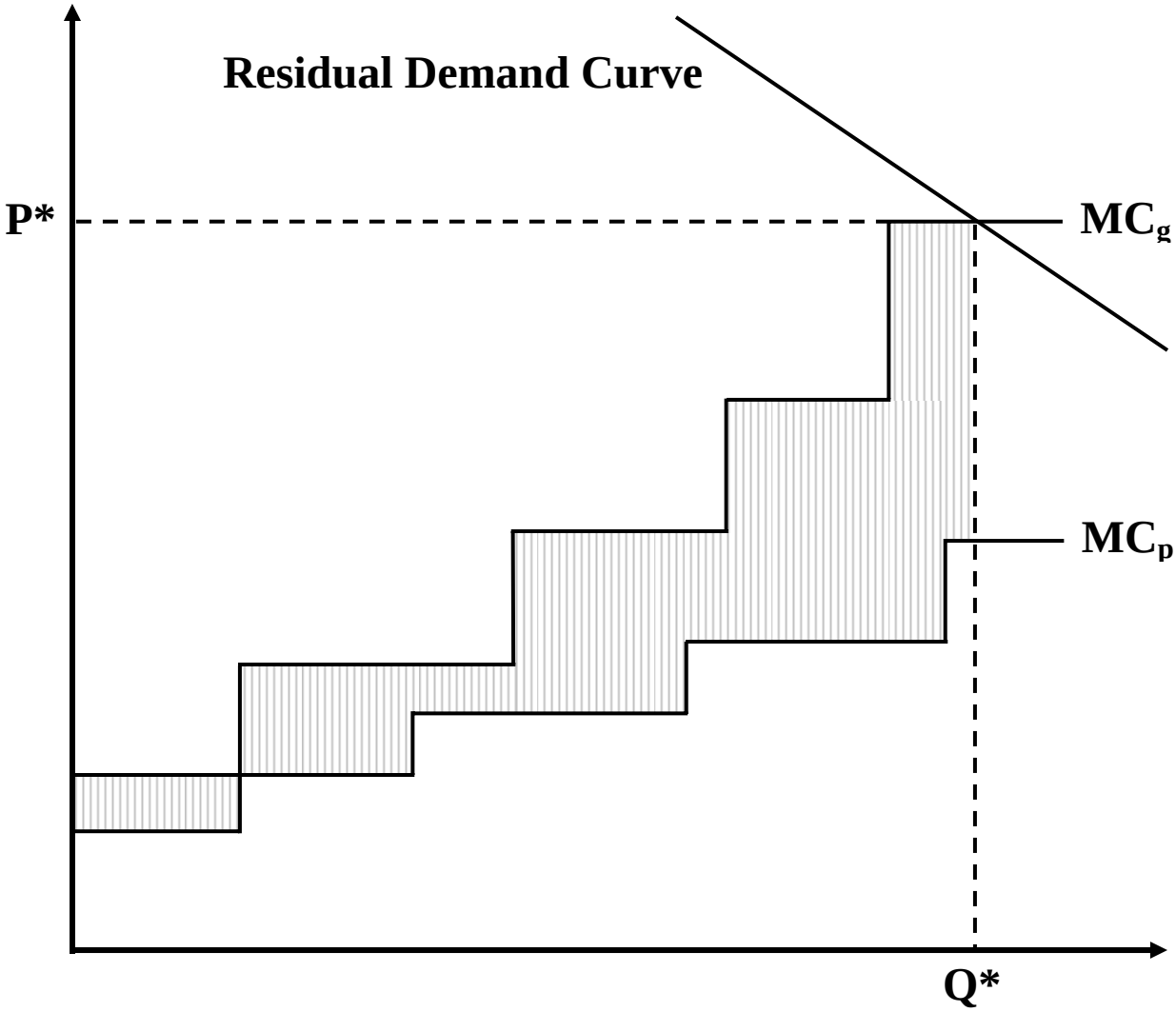

Figure 10: The Impact of Capacity Divestiture on a Pivotal Supplier

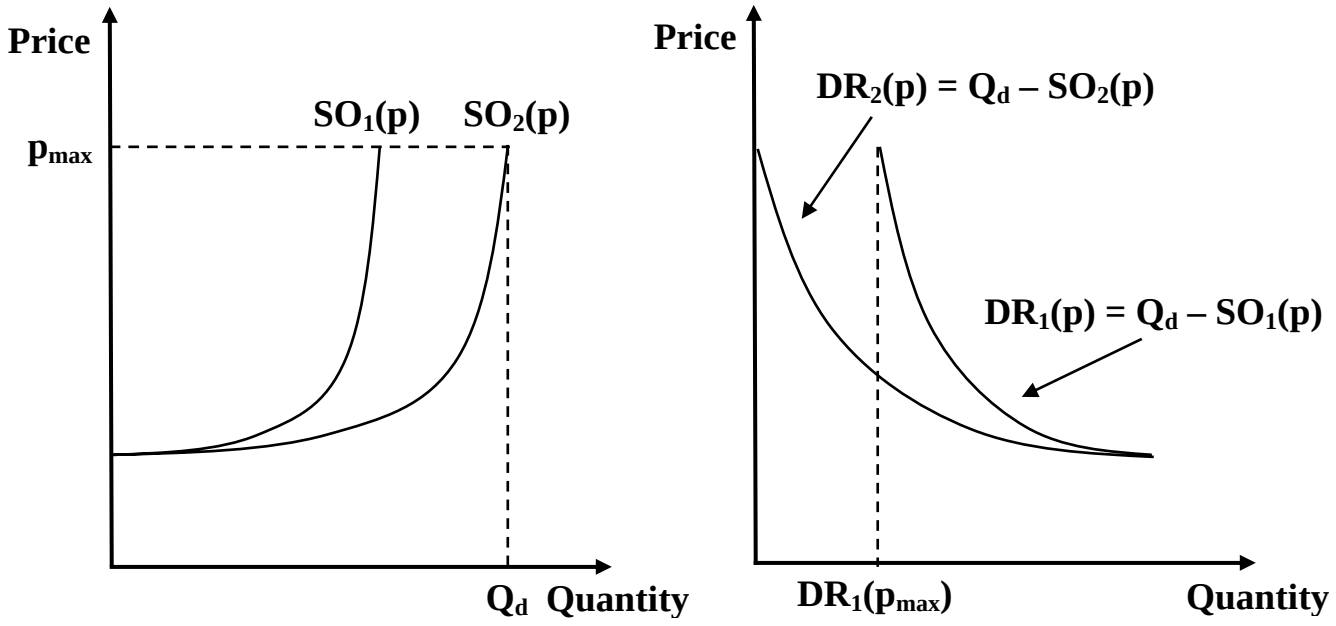

Figure 11: Residual Demand Elasticity and Price-Responsive Demand