Regulating Competition in Wholesale Electricity Supply

by

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

September 5, 2005

Very Rough Draft

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 difficult. There has been a number of much-publicized disasters–the California electricity crisis during the period June 2000 to June 2001 and the New Zealand market meltdowns during June to September of 2001 and 2003. Even those wholesale markets generally acknowledged to have ultimately delivered benefits to consumers in countries such as the United Kingdom and Australia have experienced substantial problems with the exercise of unilateral market power by large market participants.

The experience of the past ten years suggests that, although there are opportunities for consumers to realize benefits from electricity industry re-structuring, these benefits are small relative to 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 these other two industries. This conclusion motivates three major questions which will be addressed in this chapter. First, why has the experience with electricity structuring been so disappointing, particularly in the United States? 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 United States and other industrialized countries?

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. Because the technology of electricity transmission and local distribution overwhelmingly favors a single network for a given geographic area, a regulatory process is necessary to set network access prices and to monitor the operation and expansion of the transmission network. However, the mechanism used to allocate use of the transmission network among market participants and the configuration of the transmission network exert a dramatic impact on short-term wholesale market outcomes. In addition, the mechanism used to determine the location and magnitude of expansions to the transmission network has an enormous impact on the magnitude and location of new generation investments. Different from virtually all other industries, a restructured electricity supply industry requires certain segments to be regulated and how these segments are regulated significantly impacts market outcomes. Consequently, the regulatory process in a restructured electricity supply industry 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. This requires a dramatic change in the nature of the regulatory process relative to the one that existed under the formervertically-integrated utility regime.

This chapter describes the major differences in the challenges facing the regulatory process in these two regimes. In the former 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. This chapter will describe 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. 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. 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, (2) impose penalties and sanctions on market participants for their behavior, and (3) even explicitly set the prices that market participants receive for their production. This chapter provides a theoretical framework for understanding how best to make these choices in order to design a wholesale market that benefits consumers relative to the formervertically-integrated utility regime. 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.

Examples from wholesale markets in industrialized and developing countries will be used to illustrate the importance of effectively addressing each aspect of the market design process. Although there is no single optimal wholesale market design, all successful market designs appear to have adequately addressed each of these dimensions of the market design process in a manner suited to the initial conditions in the industry and political constraints the re-structuring process faced when the wholesale market was formed. This point is illustrated through examples from industrialized countries and developing countries where different economic and political constraints on the industry in each country led to different market design choices. The divergent market design choices across these countries arguably improved the performance of the wholesale market in that country, because each choice was tailored to the economic and political constraints faced by that country.

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 to a number of ways to increase the likelihood of a restructuring process in the US that benefits consumers.

2. History of Electricity Supply Industry and Its Regulatory Oversight

This section briefly reviews the technology of electricity producing and delivering electricity. This is followed by a discussion of the major features of the formervertically-integrated utility regime and regulatory process that governed it. Because there are a number of excellent comprehensive discussions of the evolution of electricity supply industry and the regulatory oversight process, I focus on the features of the production technology, industry structure and regulatory oversight that highlight the important dimensions of the difference between the restructured regime and vertically-integrated, regulated monopoly regime.[1]

---

[1]Joskow and Schmalensee (1983) provide a comprehensive discussion of the technology of electricity supply and the historical evolution of the industry in the US.

2.1. Technology and Market Structure in Electricity Supply Industry

The electricity supply industry is typically divided into four stages: (1) generation, (2) transmission, (3) distribution, and (4) retailing. Generation is the process 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 in 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 and selling wholesale electricity 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. These firms owned and operated generation assets, the transmission network, and local distribution network required to deliver electricity throughout their geographic service area.

There is some debate surrounding the rationale underlying the origins of this industry structure. The conventional view is there were economies to scale in the generation of electricity at the level of demand served by most electricity utilities. There were also thought to be significant economies to scope between transmission and generation at the level of demand and size of the geographic region served by most vertically-integrated utilities. Both of these industry characteristics argue in favor of a single vertically-integrated supplier for a given geographic area, which implies the need for a regulatory process to set the price the monopoly supplier is allowed to charge and the terms and conditions under which it can charge this price.

An alternative rationale for the vertically-integrated, regulated monopoly market structure has been put forth by Jarrell (1978). He argues that this market structure arose from the early years of the industry when these local monopolies found it extremely 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.

Although there are a number of municipally-owned, vertically-integrated utilities and an even larger number of customer-owned electricity cooperatives serving rural areas of the US, until the re-structuring process began in the late 1990s, the vast majority of US consumers were served by privately-owned vertically-integrated utilities. As noted in Joskow (1974), customers served these 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. Joskow (1989) provides an extremely insightful 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.

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 turned out to be extremely costly to consumers. This led to a general dissatisfaction with the vertically-integrated utility regulatory paradigm. 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, the England and Wales was the first to embark on this process, and a number other industrialized countries quickly followed. 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)[2]. 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. All of these factors

combined to provide the impetus for the formation of 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]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.

2.2. Regulatory Process Governing Vertically-Integrated Utility Regime

The regulatory process in the US is complicated by the fact that the federal government has jurisdiction over all interstate commerce and state governments have jurisdiction over all intrastate commerce. The technology of electricity transmission makes the distinction between interstate and intrastate sales of electricity somewhat arbitrary. Specifically, electricity flows along the path of least resistence thorough the transmission network. Consequently, even though a supplier would like to ship more electricity from point A to point B (link AB) than from point A to point D (link AD), if the electrical resistance in the transmission network between A and B and A and D are equal then 1 megawatt-hour (MWh) injection at point A results in a 0.5 MWh of flow on link AB and a 0.5 MWh flow on link AC. Consequently, rather than treating electricity like other goods that can be physically transported from point A to C, a more realistic model for an electricity network is that suppliers inject electricity into the network at their location and loads withdraw electricity at their location in the network. The energy that is injected into the network flows according to Kirchhoff's First and Second Laws. Because electricity flows according to the laws of physics rather than where sellers or buyers would like it to flow, determining how much of the electricity consumed in one state was actually produced in another state if the two states are interconnected through the bulk transmission network is extremely difficult. As consequence, a somewhat arbitrary distinction between federal-jurisdictional and state-jurisdictional energy sales must be made for transactions that flow over the interconnected interstate transmission network if the buyer and seller are located in the same state. 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 subsequently became the Federal Energy Regulatory Commission (FERC) in 1977 to regulate wholesale energy transactions. 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." If FERC determines that they are not just and reasonable, then it has considerable discretion to take actions to make wholesale electricity 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.

Under the vertically-integrated regime, state-level regulation of retail electricity rates effectively controlled the price these utilities paid for wholesale electric power. These utilities either owned all of the generation units necessary to meet their retail load obligations or had long-term contract commitments for energy and generation capacity sufficient to meet their retail load obligations. The implicit regulatory contract between the state regulator and the utilities within its jurisdiction was 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 had an obligation to serve all demand in its geographic service area at this regulated price. Although these vertically integrated utilities sometime made short-term electricity purchases from neighboring utilities, virtually all of their retail energy obligations where met either from long-term contracts or generation capacity owned and operated by the utility.

As Joskow (1989) notes, this state-level regulatory process was far from perfect in a number of ways. 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. As Joskow (1974) describes in detail, nominal prices remained unchanged for a number of years during the 1950s and 1960s. However, during the late 1970s and early 1980s when input fossil fuel costs rose dramatically, many utilities filed to increase their prices a number of times in rapid succession. A second imperfection in the rate-making process is that regulators have considerable discretion to determine what costs are prudently incurred and therefore the utility is entitled to recover in the prices it is allowed to charge. Joskow (1989) stresses that the "used and useful" regulatory standard is used to justify whether an investment is prudent. Specifically, 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 that can allow regulators to disallow cost recovery for certain investments that seemed needed at the time they were made but subsequently turned out not to be used and useful.

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 review 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 facilities failed to materialize. The fundamentally political nature of the regulatory price setting process necessitated these reviews and cost disallowances. 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. This meant that utility shareholders had to cover the costs of these ex post "imprudent" investments in coal and nuclear generation facilities. The utility's shareholders had to cover the losses associated with these 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 significantly 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 supply 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 (2000) 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 operation in California in April of 1998.

Before closing this section, it is important to highlight two important features of the regulatory process governing wholesale electricity in the US that will play a role in our later discussion. First, FERC historically had very little role in regulating wholesale electricity prices, even as more wholesale transactions occurred. Joskow (1989) point 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). 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 public utility commissions exerted almost complete control over retail electricity prices. As we will discuss below, with the advent of wholesale electricity markets with significant participation by pure merchant suppliers–those with no regulated retail load obligations–the ability of state PUCs to control retail prices is severely limited. FERC's role in controlling retail prices is directly enhanced by the extent to which major load-serving entities in the state no longer own generation assets and by extent to which they purchase their remaining wholesale energy needs on short-term markets.

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 bilateral negotiation or through the operation of a bid-based wholesale electricity market. Because markets can and often do set prices substantially in excess of just and reasonable levels, this requirement of the Federal Power Act considerably complicates the process of regulating wholesale markets in US, increasingly so given the experience with California market during the period June 2000 to June 2001. Wolak (2005) discusses the process used by FERC to determine whether to allow suppliers to sell at market-determined prices, rather than cost-of-service prices. 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 Regulatory Oversight

This section describes the technological characteristics of electricity supply and the political and economic constraints facing the industry that make wholesale electricity markets so susceptible to the exercise of unilateral market power. This includes a discussion of the new institutions and market participants in a wholesale electricity market that can enhance the ability of suppliers to exercise unilateral market power. These entities create substantial challenges, discussed in this section, for industry oversight that extend beyond the traditional boundaries of the industry-specific regulator.

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 the one that occurred in the eastern US and Canada on August 13, 2003. The entity that operates the transmission network for a given geographic region, typically called in the independent system operator (ISO) in the wholesale market regime, reserves a certain amount of generation capacity for what is called automatic generation control (AGC) to provide second-to-second system balance services. A generation unit on AGC is directly connected to computers located at the ISO and responds automatically to second-to-second dispatch instructions issued by the ISO's system operation software. This second-to-second system balance service is typically called regulation reserve, because during most hours of the day unit providing regulation reserve do not provide any net energy to the transmission network beyond the amount the unit is pre-scheduled to operate. Units providing AGC are intended to only move their output level up and down relative to this operating point within the hour. For example, a 300 MW unit providing regulation could be scheduled to produce 200 MWh for the hour and provide 50 MW of regulation reserve. This means the unit could produce using as much as 250 MW of capacity, 50 MW above its 200 MWh schedule, for some portion of the hour and as little as 150 MW for other parts of the hours, which is 50 MW below its 200 MWh schedule.

It is very costly to store electricity. The most efficient way to store electricity is using a pumped storage facility. This involves first pumping water uphill into a reservoir and then running it through a turbine to produce electricity when it needed. To produce 2 MWh in this manner requires expending at least 3 MWh to pump the water required uphill. In addition, a specialized site that has the required water reservoirs above and below the turbine is needed. There is also the substantial cost of constructing the pumped storage facility and the going forward cost of operating it.

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 given generation unit in an hour. A generation unit with a 500 MW nameplate capacity can typically produce only slightly more (approximately 5%) than 500 MWh within a given hour. This additional output is produced at a higher variable cost and results in a greater likelihood of a forced outage from the generation unit.

Delivery of the product consumed must take place through a potentially congested transmission network. If a supplier owns a portfolio of generation units that are connected at different, but relatively nearby, 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. A simple example illustrates this point. Suppose there is large number of suppliers at location A and a single supplier at location B. The demand at location A is zero and demand at location B is 400 MW and the transmission capacity between A and B is 300 MW. If the suppliers at location B restricts output from its units to cause congestion into location B, it can become a monopolist facing a residual demand of 100 MWh, because the maximum amount that the suppliers at location A can sell is 300 MWh, the capacity of the transmission line, despite the fact that demand at location B is 400 MWh.

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 was when this KWh was consumed. Part of the reason for this single fixed retail price is the fact that most residential meters are only capable of recording the total amount of KWh consumed between consecutive meter readings. Consequently, a significant barrier to implementing retail electricity prices that reflect wholesale market conditions is the availability of metering technology that records hourly consumption for all hours of the month. With these meters in place, consumers can benefit from purchasing electricity at an hourly level. For reasons discussed in Section 8, the vast majority of utilities that have managed to install meters that record the hourly consumption for all hours of month find it extremely difficult to obtain approval from their state PUCs to charge retail prices that vary with wholesale market conditions.

The technology of 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. The longer is the time lag between negotiation of the agreement to sell and the actual delivery of the electricity sold, the more suppliers are able to compete to provide the 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 the market. 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 product. 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, it is clear that the least competitive market is the real-time energy market. Only suppliers operating their units in real-time with unloaded capacity or quick-start combustion turbine at locations in the transmission network that can actually supply this energy need are able to compete to provide this energy.[3]

For this reason alone, it is not surprising that real-time prices are far more volatile than day-ahead prices, which are far more volatile than month-ahead or year-ahead electricity prices. It is easy to imagine that an electricity retailer would be willing to pay $1000/MWh for 100 MWh in the real-time market, or even $5000/MWh, if that meant keeping the lights on for its retail customers. However, it is extremely hard to believe that this same load-serving entity (LSE) would pay much above the long-run average cost of producing electricity for this same 100 MWh electricity to be delivered two-years in the future.

[3]A generation unit has unloaded capacity if its instantaneous output is less than the unit's maximum instantaneous rate of output.

This logic demonstrates 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 energy 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 markets, this time horizon is between 18 months to 2 years, meaning that if system conditions arise that create opportunities for suppliers 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. Although the opportunities to exercise system-wide market power are relatively short-lived, the experience of a number of wholesale markets has demonstrated that suppliers with unilateral market power are able to raise market prices substantially during this time period, and thereby inflict substantial harm to consumers.

Electricity suppliers possess differing degrees of system-wide 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.

These two determinants of market power imply 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 sufficient competition to discipline its bidding behavior into the wholesale market.[4] 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 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 not even the LSE for 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 owner to bid whatever the market will bear for any energy these units provide. This incentive exists regardless of the locational pricing scheme used by the wholesale market operator.

---

[4]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.

This point deserves emphasis: Absent a local market power mitigation mechanism, the bid of the unit or units with local market power must be taken before lower-priced bids from other firms. The configuration of the transmission network and location of demand makes this unit the only one physically capable of meeting the energy need. Without some form of regulatory intervention, these suppliers will be paid at least their bid price to provide the needed electricity willingly. The configuration of the existing 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.

4. Regulatory Challenges in Wholesale Market Regime

A wholesale electricity market requires substantially more sophistication and economic expertise from the regulatory process at both the federal and state levels than it did under thevertically-integrated utility regime. Errors by the regulatory process in the formervertically-integrated utility regime typically causes a relatively slow rate of financial harm to consumers and producers. Regulatory errors in the wholesale market regime can create substantial opportunities for suppliers to exercise unilateral market power and cause enormous wealth transfers between producers and consumers in a very short time period.

The regulatory process must set output prices and determine the prudency of investment decisions for the transmission and distribution network owners, tasks it formerly carried out for the vertically-integrated utility. It must now take on the additional task of designing and monitoring the wholesale and retail market segments of the industry. The combined federal and state regulatory process must determine what wholesale and retail market rules will result in wholesale prices that allow suppliers and retailers the opportunity to recover their prudently incurred costs. It must also decide when and how to intervene when market mechanisms cannot be relied upon to set prices because of market power concerns. The existence of a wholesale market results in the entry of a number of new entities and market participants that require regulatory oversight. The remainder of this section details the unique challenges faced by the regulatory process in a wholesale market regime.

The uncertain availability of generation units and portions of the transmission network because of forced outages implies that system conditions can arise when virtually any generation unit owner in the wholesale market possesses substantial local market power. Fortunately, these system conditions are often short-lived because the transmission line or generation unit is repaired. Consequently, a prospective local market power mitigation (LMPM) mechanism that anticipates these circumstances and provides bid mitigation when they arise is a necessary component of any wholesale market design. The wholesale market regulatory process must design and implement an effective LMPM mechanism or system conditions will increasingly occur where suppliers have the opportunity exercise substantial local market power.

Another unique feature of bid-based wholesale electricity markets is that market design flaws that cause little harm during most system conditions lead to substantial consumer harm under certain system conditions. The experience of the California market illustrates this point. From its start in April 1998 until April 2000, the California market was probably the most competitive wholesale market in the US. Average wholesale prices over this period were less than $35/MWh and the average hourly magnitude of the market inefficiencies as measured by the methodology given in Borenstein, Bushnell and Wolak (2002), hereafter BBW, was less than or very close to equal to those measured by Mansur (2003) for the PJM market and Bushnell and Saravia (2003) for the New England market. This level of market performance 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.

The amount of hydroelectric energy available from the Pacific Northwest during the summer of 2000 was significantly less than the previous two summers. Wolak (2003a) shows that this event led to the five largest fossil fuel electricity suppliers in California facing significantly less elastic residual demand curves than they did during the first two summers the wholesale market operated. As a consequence, these suppliers found it in their unilateral interest to bid less aggressively into the spot market in order to raise wholesale electricity prices in California. As discussed in Wolak (2003b), this strategy was not unilaterally profitable during the first two years of the market because the greater availability of hydroelectric energy from the Pacific Northwest and inexpensive coal-fired energy from the Desert Southwest during the summer months caused these suppliers to face significantly more elastic residual demand curves.

This change in competitive conditions during the summer of 2000 enabled in-state suppliers to raise prices substantially through their unilateral actions. For example, during the summer months of June to September of 2000, the average difference between the actual price and the BBW (2002) competitive benchmark price was more than $70/MWh, which is more than twice the average price of electricity during the first two years of the market. The California experience demonstrates that some market design flaws, in this case insufficient forward contract coverage of final demand 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 cause enormous harm to consumers. Effective wholesale market regulatory oversight is necessary to intervene as quickly as possible to limit the potential damage when these system conditions arise.

While it is not possible to rule out coordinated actions among the major electricity suppliers in the Western Electricity Coordinating Council (WECC) as a potential explanation for the enormous increase in wholesale prices over the period June 2000 to June 2001, as shown in Wolak (2003a), this is not necessary. The behavior of prices during this time period relative to the first two years of operation of the market can be explained by the unilateral expected profit-maximizing actions of the major California fossil fuel suppliers given the distribution of residual demand curves they faced. 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.

From the perspective of antitrust law, the most surprising aspect of the period June 2000 to June 2001 in the California market is that despite estimated total market inefficiencies of close to $20 billion, virtually all of which was due to the exercise of unilateral market power, US antitrust law did little to prevent this enormous wealth transfer from occurring. Moreover, following the Enron bankruptcy and disclosure by a number of energy trading firms (in very unflattering terms) that they did attempt to exercise all available unilateral market power in the California market, US antitrust law has been unable to obtain refunds of any of these market power profits. This outcome has occurred because US antitrust law does not prohibit firms from fully exploiting their unilateral market power. Because of the enormous potential harm from the exercise of unilateral market power in electricity, other regulatory safeguards are necessary.

Besides the need to intervene to correct market design flaws after they are determined to be harmful, there is also a need to engage in prospective market monitoring to find market design flaws that lead to substantial harm by less noticeable means. For example, certain aspects of the wholesale market design can increase the likelihood that coordinated actions to raise prices might occur. Aspects of the market design can also enhance the ability of suppliers to exercise their unilateral market power. This logic suggests that another important role for the regulatory process is to monitor the wholesale market 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. Particularly, during the initial stages of the wholesale market regime, this prospective approach to regulatory oversight should implement market rule changes before they expose consumers to significant harm.

This role for the regulatory process also has a pedagogical component. The transition to a wholesale market regime involves a dramatic change in behavior by a number of market participants. Companies that fail to adapt to the new regime are very likely to go bankrupt and exit the industry, but there are often significant external costs to consumers associated with this outcome. Consequently, the regulatory process should take prospective actions to encourage adaptation to the new regime and limit the resulting external costs if this change in market participant behavior does not occur. An example of a necessary change in behavior is the need for retailers to hedge spot price risk, something that was unnecessary in the former vertically-integrated regime because electricity retailers typically owned enough generation capacity to meet their load obligations.

A significant regulatory challenge in a number of markets around the world is how to provide strong incentives for retailers to engage in the efficient amount of forward contracting, while at the same time not exposing consumers to sustained periods of very high spot prices. A purely market-based solution would be to expose retailers to the risk of bankruptcy by having extremely high price caps or bid caps on the spot market. The fear of bankruptcy associated with a sustained period of extremely high spot prices when the retailer has a significant spot market exposure will cause retailers to engage in the appropriate amount of hedging of spot price risk. This strategy has worked remarkably well in Australia, where the bid cap on the spot market was $5,000/MWh, and has recently been raised to $10,000/MWh. Spot prices at or near these levels occasionally occur, but because of the high levels of forward contracting by Australian retailers these prices do not cause significant harm to consumers.

In the US and New Zealand there have been a number of bankruptcies of large retailers as a result of a sustained period of very high spot prices that occurred when the retailer had a significant exposure to the spot market. Over-reliance on the spot market by retailers in the US is not surprising given the decision of the Federal Energy Regulatory Commission (FERC), the US wholesale market regulator, to impose hour-by-hour market power mitigation in the form of relatively low bid or price caps on the spot market and market power mitigation procedures that set very low bid caps for specific generation units. This mitigation limits the volatility and level of spot prices, which dulls the incentive for retailers to sign forward contracts or engage in other hedging arrangements to limit their exposure to the spot market. Consequently, as we discuss in Section 8 regulatory invention to limit the volatility of spot prices creates the need for regulatory intervention to mandate that retailers purchase a portfolio of financial contracts to hedge their spot price risk.

Particularly for the US, where retail market regulation is the domain of the state regulatory commissions and wholesale market regulation is the domain of FERC, this strategy by FERC may be necessary to gain the consent of the state regulator to relax its control over retail electricity prices. Forward contracting requirements on electricity retailers are a part of virtually all Latin American markets. The regulatory process typically mandates forward contracting requirements for retailers at various time horizons in advance of delivery to ensure there is adequate generation capacity to serve demand and that no retailers are overly exposed to the spot market.

Regulating forward contracting levels has an additional spot market competitiveness benefit. If these purchases are structured as fixed-price forward contracts for a fixed amount of energy in the hour, they have very beneficial impacts on the competitiveness of shorter-term energy markets. As discussed in Wolak (2000a), forward contract obligations by a supplier make it unilateral profit-maximizing to bid more aggressively in the spot market. Moreover, Wolak (2000a) also emphasizes that forward contract obligations by one supplier can make it unilaterally profit-maximizing for other suppliers to bid more aggressively, regardless of their own forward contract holdings. The most successful wholesale electricity markets, as judged by the competitiveness of their spot markets, are those with where only a very small fraction of the total amount of electricity consumed is actually purchased in the spot market.

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 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. 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. The industry-specific regulator is best-suited to develop the expertise necessary to determine the transmission network that maximizes the competitiveness of the wholesale electricity market.

The ISO 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 and 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 system operation services and for that reason regulatory oversight is needed to ensure that it is operating the grid in a least cost manner to benefit market participants rather than the management and staff of the ISO. In virtually all markets in the US, 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.

Traders 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. For example, a supplier may sell its output in a fixed-price forward contract to an energy trader in order to obtain price certainty for some amount of the expected output of its generation units. The energy trader may then sell this forward contract to a retailer, hopefully at a higher price, so that the retailer can obtain wholesale price certainty for some fraction of its retail load obligations.

Arbitrage is another trader activity that has created considerable controversy among politicians and the press. Arbitrage is an attempt 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, arbitrage involves buying the right to inject electricity at one location and selling the right to inject electricity at another location. This is often incorrectly described as buying electricity at one location and selling it at another location. However, as discussed above, it is not possible to take possession of electricity and transport it from one location to another. Consequently, selling an injection of electricity at location A and buying an injection at location B is taking a gamble on direction and magnitude of congestion between these two locations in the transmission network. If there turns out to be no congestion between the two locations then the trader can fulfill his obligation to inject at location A by purchasing electricity at the real-time price at that location and his obligation to sell an injection by selling energy at the real-time price at location B. In this case, the trader's profit is the difference between the day-ahead prices at locations A and B.

Virtually all arbitrage 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 location A and the real-time price turns out to be higher than day-ahead price at location A, then the trader must fulfill the commitment to provide 1 MWh at location A by purchasing at the higher real-time price. This arbitrage 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 and market design flaws and increasing the economic harm they cause. Although virtually all of the Enron trading strategies described in three memos released by FERC in the Spring of 2002 could be classified as risky arbitrage strategies that had the potential to enhance market efficiency, but a few appeared to degrade system reliability or market efficiency only. 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.

5. Responsibilities of Regulatory Process

This section states the primary goal of the industry regulatory process and lays out the details of the four major responsibilities of the industry-specific regulator.

5. 1. Goal of Regulatory Process

Because electricity markets are so susceptible to the exercise of unilateral market power, the primary goal of the industry-specific regulatory process is to limit the ability of market participants to engage in behavior that degrades system reliability and market efficiency, the two major adverse consequences of the exercise of market power.

It is impossible to prevent firms from exercising unilateral market power. This would imply the existence of a perfect regulatory process. The market or system operator would need to know each supplier's minimum cost of producing power. It could then dispatch suppliers based on their minimum cost of producing power. However, if such a regulatory process existed there would be little need to introduce a competitive market because, by assumption, a lower average cost of supplying power to consumers could be achieved by paying suppliers only their minimum cost of production, rather than the market-clearing price set through a process where all suppliers bid to maximize their expected profits for all of the energy they produce. Consequently, any mechanism used to mitigate market power is necessarily imperfect in the sense of being unable to protect consumers from the exercise of all the market power a supplier possesses.

By this same logic, there are no perfectly competitive markets. However, there are many markets that yield outcomes very close to those predicted by the perfectly competitive ideal.

Unfortunately, electricity is not always one of these markets. Consequently, the market designer is faced with the choice between an imperfectly competitive market and an imperfect regulatory mechanism to determine the compensation paid to a market participant. A social welfare maximizing market designer would make this choice based on which mechanism entails the smallest loss in social welfare.

This logic implies a regulatory process that provides incentives for efficient market outcomes, instead of focusing on preventing firms from exercising all unilateral market power. The regulatory process should provide the strongest possible incentives for least-cost provision of electricity to final consumers consistent with the long-term financial viability of the industry. Running a regulatory process is costly and regulatory invention even more so. Therefore, it is important to account for these costs in the design and operation of the regulatory process. Specifically, the regulatory process should first focus on actions that have a very high market efficiency benefits relative to their implementation costs. The regulator should also periodically review the costs and benefits all of aspects of the regulatory process.

5.2. Three Major Responsibilities of Regulatory Process

I will now describe the three major responsibilities of the industry-specific regulator and how these should be carried out. They are: (1) disseminating information to existing and prospective market participants, (2) ensuring compliance with all the market rules, and (3) protecting against behavior that degrades market efficiency and system reliability. Successfully fulfilling each role requires much greater regulatory authority and sophistication on the part of the regulatory process than the previous one.

5.2.1. Smart "Sunshine Regulation"

A minimal requirement of any regulatory process is to provide "intelligent 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. Specifically, all data submitted to real-time market and produced by the system operator should be immediately released to the public. Little net energy sales should take place through the real-time market, because it is operated primarily for reliability reasons and all market participants have a common interest in the reliability of the transmission network. Immediate data release best serves these reliability needs.

There should be no limitation on the regulator's access to data either submitted to the system operator by market participants or produced by the system operator. Besides all of the information needed to operate the energy and ancillary services markets and the transmission network, the regulator should also have the ability to request information from market participants on a confidential basis to perform further analyses. Rather than have an ex ante limitation on the type of data it can request, the regulator should have open-ended authority to request information subject to an economic cost-benefit test. To enforce this authority, the regulator should also have the ability to impose financial penalties on market participants that fail to provide the requested data in a reasonable period of time.

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). On the other hand, 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 public data release should identify the market participant and specific generation unit associated with each bid, generation schedule, or output level. Masking the identity of the market participants, as is done in all US wholesale markets, limits the disciplining value of public data release on 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 their behavior to the public. 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, virtually eliminates much of the potential 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 problem with 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.

For the case of the UK, the MALC anticipated punishing those market participants that exercised a significant amount unilateral market power. However, one difficulty with this approach is that the major beneficiaries of the unilateral exercise of market power are the firms that exercised little if any unilateral market power. One could therefore imagine some firms finding ways to compensate larger firms for exercising their unilateral market power so that all firms can reap benefits. A second difficulty is distinguishing the exercise of significant market power worthy of punishment from expected profit-maximizing behavior. 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 third-parties to undertake analyses of market performance. The US policies on data release severely 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. For example, some 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.

The 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. 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, 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. Although 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, 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 difference competitive benchmark levels as described in BBW (2002). Keeping this real-time data confidential prevents this potentially important form of public scrutiny of market performance from occurring.

This data can also be used to undertake third-party studies of whether coordinated actions, explicit or tacit, are occurring. Although these third parties would find it difficult to produce evidence of explicit or tacit collusion that would lead to a conviction in a court of law, they only need to present evidence that has a high likelihood of yielding a conviction in the court of public opinion. The prospect of such adverse publicity is very likely to increase the perceived cost to market participants of engaging in explicit or tacit co-ordinated behavior to raise market prices. Despite a prohibition against conscious parallelism in under US antitrust law, I am not aware of a successfully prosecuted lawsuit against such behavior in any industry in the US. Consequently, the fear of a conviction in the court of public opinion appears to be best way to prevent such behavior.

Economic theory provides no clear predictions about the relationship between the information made available to market participants and their ability to implement less competitive market outcomes. A number of theoretical papers have identified circumstances when asymmetric information between market participants can be a very effective device for implementing less competitive market outcomes, through either unilateral or co-ordinated actions. I am also not aware of any systematic empirical evidence industries demonstrating that making more information available to market participants leads to less competitive market outcomes.

Coherent arguments in favor of masking the identity of market participants in the publicly released bid, schedule and production data are more difficult to find. Assuming that the concerns with public data release enhancing the ability of market participants to coordinate actions had been addressed, it is difficult to determine what market efficiency-enhancing benefit results from masking the identity market participants. Masking the identity of the market participant only limits the "sunshine regulation" value of public data release.

An important aspect of the public data release question is the distinction between data that the regulator can request and receive from market participants and data that must be released to the public. There is a natural boundary between these two types of data. Any data that the system operator must request from market participants or must produce in order to operate the real-time market should be released to the public. As noted earlier, the real-time market is operated primarily for reliability reasons and this data release policy is consistent with the goal of preventing harm to system reliability and market efficiency.

Public release of any information on forward market positions or transactions prices, where the large volumes of energy are typically traded, does not serve this goal. This is information about a market participant that is unnecessary to operate the real-time market, although it does impact the bidding, scheduling or production behavior of that market participant, as discussed in Wolak (2000a). Knowledge of these financial positions is not needed by the system operator to run the spot market or the transmission network.

Because of the fundamentally financial nature of forward market transactions sold by electricity suppliers, it is very difficult to get accurate information on the true forward market positions of electricity suppliers. They can re-trade forward market obligations among themselves to yield forward market positions far above or below their expected production of electricity. For this reason, even if the regulator attempted to collect this forward market data from suppliers on a regular basis it would not be very useful. For example, if the regulator specified a minimum quantity of forward contract sales for each supplier it regulated, these suppliers could undertake forward contract transactions with affiliates not subject to regulatory oversight to meet these minimums. Moreover, those affiliates not subject to oversight by the regulator could then re-construct their holding company's desired forward contract position. Consequently, routinely collecting the forward contract positions of suppliers could cause them to render this information of little or no use to the regulator through affiliate transactions.

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 spot 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.

Before the start of each year, suppliers should submit to the system operator a schedule of planned outages for each of their units. The system operator would then compile the planned outage schedules submitted by all suppliers and verify that they do not compromise system reliability. If they do, then the system operator will suggest modifications to achieve a schedule of planned outages for all units consistent with reliable network operation on annual basis. Although the system operator should attempt to accommodate the wishes of each supplier, it must have the ultimate authority to set the final schedule for all planned outages. Once this planned outage schedule is set, it should be released to the public. Modifications of these unit-level planned outages schedules during the year are subject to the approval of the system operator. These modifications should be released to the public once they are approved.

A similar process should be followed for scheduling planned transmission line outages. The system operator should coordinate the planned transmission outage process with all of the transmission owners and the generation unit owners. It should also make the final decision on when both generation units and transmission lines can be taken out for maintenance.

To limit the incentive for “sick day” unplanned generation outages, the system operator should 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. Such a unit should be assumed to be forced out, because the owner is not offering this capacity to the market. The system operator should 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 should 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 spot 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.

5.2.2. Ensuring Compliance with Market Rules

Many market outcomes that are harmful to system reliability and market efficiency could be prevented if all market participants fulfilled their contractual obligations. Because the actions of each market participant impacts system reliability and market efficiency, this implies that all parties have a common interest in honoring their contractual obligations. If the cost of violating a contractual commitment or market rule is less than the unilateral benefit from this action, the market participant will find it profitable to violate, which also adversely impacts system reliability and market efficiency. This logic implies that the second responsibility of the regulatory process is to: (1) design market rules to resemble publicly verifiable contractual obligations, and (2) determine the appropriate penalties and sanctions to deter violations of these rules without adversely impacting market efficiency or system reliability.

Not all market rules are defined to resemble publicly verifiable contractual obligations. Prohibitions against market manipulation or the abuse of market power are prime examples. These prohibitions have done very little to prevent harmful market outcomes. The California market rules contained these prohibitions, but they did little to prevent the events of June 2000 to June 2001. The third responsibility of the regulatory process is to prevent harmful market outcomes that cannot be prevented by market participants obeying market rules that resemble publicly verifiable contractual obligations.

A large fraction of harmful market outcomes can be prevented and the costs of operating the market and the costs of participating in the market will be lower if all market participants are confident that all contractual commitments will be honored regardless of system conditions. Contract enforcement costs stem from ambiguous or overly broad market rules or market rules that are not, or cannot be, enforced. A transparent rule that can be rigorously enforced is superior to an overly broad rule that is difficult to enforce. Irregular enforcement, either because of imprecise rules or inconsistent effort, increases the cost of market participation. This can also lead to increased market rule violations as more market participants push the boundaries of acceptable behavior.

This logic implies that the regulator should divide market rules into two categories: (1) those that resemble publicly verifiable contractual obligations with little subjective judgement to determine compliance, and (2) those that require a formal administrative process to determine compliance. Rules in first category should be written to limit ambiguity and simplify enforcement. Those in the second category should have pre-specified administrative processes that deters behavior harmful to system reliability and market efficiency because of the large amount of judgement associated with determining that a violation has occurred.

The first type of market rule should be written so that a violation resembles the process of issuing a speeding ticket as closely as possible. If the regulator measures the speed of the car using a publicly verifiable measuring device and finds that the car's speed exceeds the posted limit, then the regulator should assess a pre-specified penalty. The penalties and sanctions process should not involve a finding of intent in order for the regulator to assess a penalty. An example of a market rule violation covered by this mechanism is a failure to comply with terms implied by a bid into the wholesale market. One example is a market participant submitting a bid to supply a fixed quantity of energy within a given response time and then failing to meet this commitment. Suppose the supplier bids to provide 50 MWh of energy in 10 minutes from the time the bid is accepted. If the supplier fails to provide any of the purchased energy when it is called upon, the unit owner should be penalized for failing to meet this contractual commitment.

Both types of market rules require penalty and sanction mechanisms, but for slightly different purposes. In both cases, penalties and sanctions are imposed to deter market rule violations. For the market rules where determining compliance is straightforward, the penalties and sanctions are the primary mechanism for deterring violations. For case where subjective judgement is required to determine a violation, the penalties and sanctions are the ultimate backstop, but the administrative process is the primary mechanism for preventing harmful market outcomes.

5.2.3. 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. The regulator must also determine when a market rule detracts from system reliability and market efficiency and suggest and implement the necessary changes in this market rule. The regulator must determine when market outcomes cause 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.

Local Market Power Mitigation (LMPM) Mechanism. 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. 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. It is increasingly clear to regulators around the world, particularly those that operate markets using Locational Marginal Pricing (LMP), that formal regulatory mechanisms are necessary to deal with the problem of insufficient competition to serve certain local energy needs.

Formulate and Implement Efficiency-Enhancing Market Rule Changes. One lesson from the past decade of electricity supply industry reforms around the world is that it is easy to make extremely costly mistakes and very difficult to avoid making any mistakes. The only thing a government or region considering reform of its electricity supply industry can be sure of is that mistakes will be made in the initial market design and that the regulatory institution will face a number of difficult challenges. The enormous number market rule changes made in all of the US wholesale markets since they each began operation less than eight years ago attest to the process of continual improvement involved in regulating a wholesale electricity market. The regulator process should therefore have a strong market monitoring component which continually analyzes market outcome data to detect small market design flaws before they cause significant harm. The regulator must determine which market rules detract from market efficiency or system reliability and formulate and implement the appropriate market rule changes. Because the level and geographic distribution of demand, the mix of input fuels used and ownership shares for generation capacity in the control area, and the configuration of the transmission network can all change over time, market rules must also change. The regulator must continually analyze and assess the market efficiency impacts of all market rules. Once it has identified a deficient market rule, the regulator must then work with the system and market operators to devise the necessary remedy. This duty underscores the need for the regulator to analyze market performance using the data it has compiled.

Penalize Behavior Harmful to System Reliability and Market Efficiency. 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 should be 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 market rules should contain a general provision prohibiting persistent behavior detrimental to system reliability and market efficiency. The goal of this provision is to establish a process for the regulator to intervene to prevent a market meltdown. As shown in Wolak (2003a), 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. This provision protects against the harmful exercise of unilateral market power, which is distinct from the exercise of unilateral market power, which is equivalent to maximizing profits

Determine When Market Activities Can Be Temporarily Suspended. The regulator must have the ability to suspend market operations on a temporary basis when system conditions warrant it. The suspension of market operations should only occur after 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 impose enormous harm over a very short period time. For example, during the early stages of the New England market, there were short-lived, but severe distortions in the Installed Capacity and Operating Capacity markets that eventually lead to a suspension of market activities. During the California market's first summer, one of the reserve capacity markets experienced extremely high prices for a short period. During the initial stages of the wholesale market in the state of Queensland in the Australia, unilateral market power problems became so severe that it was necessary to suspend market operations until sufficient interconnection capacity with neighboring states and generation capacity within the state could be built. 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.

6. Market Design Process

This section provide a theoretical framework for describing the important features of the market design process. The market design process is first-described in general terms using a simple 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 designer's objectives. This section then introduces the concept of a residual demand curve to summarize the constraints imposed on each market participant by the market rules and uses it to illustrate the important dimensions of the market design process.

6.1. The 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 in a country, 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 cause the simultaneous actions of all suppliers and retailers to set the lowest possible retail price consistent with the long-term financial viability of the industry. Consequently, the goal of the market design process is to devise mechanisms for compensating market participants for their actions, so that final consumers pay the lowest possible retail prices consistent with the long-term financial viability of the industry. This 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. The first is that once the market designer selects the rules for determining the revenues each market participant 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.

6.1.1. The Principal-Agent Problem

To make these features of the market design problem more concrete, it is useful to consider a very 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 a 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. In general, x is written as x(a,s) to denote 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 basic versions of this 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 $x^* = x(a^*,s)$. 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 payoff function subject to constraint the lawyer will take actions to maximize his 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 value of the owner's assets subject to the constraint that the firm manager will take actions to maximize her payoff given the scheme is in place and the fact that it must provide a higher payoff to the manager than she could receive elsewhere.

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

An example of this principal-agent model relevant to electricity industry re-structuring is the regulator-utility interaction. In this case, the regulator designs a scheme for compensating thevertically-integrated utility for the actions that it takes recognizing the fact the once this regulatory mechanism is put in place the utility will attempt to maximize its payoff function given 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 participating in a competitive market have a 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 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 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.

In the wholesale market regime, the extent of government participation in the industry creates an additional source of differences in industry outcomes. Because the nature of the principal-agent relationship between the firm's owner and its management is different under private ownership versus government ownership, 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.

6.1.3. 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.

6.1.3.1. Individual Rationality Constraint Under a Market Mechanism

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 any amount of output at minimum cost. In other words, the firm will produce in a technically and allocatively efficient manner. As discussed above, 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 1 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 1(a) plots the inelastic aggregate demand curve and the upward sloping supply of all other firms besides the one under consideration. Figure 1(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 1(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^*:

$$(P^* - MC(Q^*))/P^* = - 1/\epsilon, \tag{1}$$

where $\epsilon = DR'(P^*)*(P^*/DR(P^*))$. Because the slope of the firm's residual demand 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.

Figure 1(a)-(b) illustrates the essential difference between the firm's unilateral profit-maximizing level of output when it has the ability to influence the market price through it own actions and its profit-maximizing level output when the supplier believes it has no ability to influence the market price. The supplier withholds output from the market because it knows that by doing so, it raises the price that is receives for all of the units it does sell.

A firm that influences market prices as shown in Figure 1(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.

Figure 1(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 producing at the point that the marginal revenue curve associated with DR(p), crosses its marginal cost curve. The only difference is that this marginal revenue curve is also equal to its average revenue curve, because DR(p) is infinitely price elastic, meaning that it is a horizontal line. Even in this extreme case, the firm continues to set prices that satisfy equation (1). However, because the slope of the firm's residual demand curve is infinite, $1/\epsilon$, is equal to zero so that equation (1) implies producing at the point that price equals marginal cost.

Figure 1 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 an infinitely elastic residual demand curve, 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 the firm possesses no unilateral market power, which means that it is unable to raise market price by these actions.

6.1.3.2. Individual Rationality Constraints under Regulation

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. Consequently, the firm must take 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. For example, 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 in setting its output price is greater than the actual price the regulated firm pays for capital services. This implies that a profit-maximizing firm facing the pricing-setting 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 different price for capital services 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 major reason 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 the opportunity to subject to the regulator to judicial review of any decision by the regulator about the level of the firm's price. 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 this decision. This aspect of the regulatory process reduces the incentive the firm has to produce its output in a least cost manner.

Wolak (1994) studies the regulator-utility interaction between California water utilities and the California Public Utilities Commission. 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 asymmetry between firm and the regulator about one aspect of the firm's production process. Even for the very straightforward 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. This result suggests that the 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 much greater.

The market designer does not need to worry about the impact of informational asymmetries in 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 regulated 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, this price is set to only recover the firm's prudently incurred costs, which can be significantly above least costs.

Consequently, the advantage of regulation is that the market price should not deviate significantly from 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, but unless the market is competitive, little incentive to pass-on these low 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 market mechanisms and regulatory mechanisms the market designer must make a choice between two imperfect worlds. Which mechanism should be selected depends on which one maximizes the market designer's objective function.

6.1.4. 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 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. However, because government-owned firm's management has little incentive to maximize profits, it also has little incentive to produce in a least-cost manner. However, 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. As discussed in the previous section, 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 1, is equivalent to saying that the firm faces a sufficiently elastic residual demand curve, this desire to raise the market price will yield market outcomes that cause significant harm to consumers.

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, however, they are in fact 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. However, for the case of the government-owned firm there is another reason why the firm is required to raise its price. That is because it is producing in a technically and allocatively inefficient manner, which is socially wasteful and therefore yields a reduced level of producer surplus relative to case of a privately-owned firm. 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.

Figure 2 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 uses twice as many units 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 2 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. However, as noted above the government-owned firm uses twice as much of society's scarce resources to produce Q^* as the privately-owned firm. Consequently, the additional benefit that society receives from having the privately-owned firm produce the good, even though it is exercising significant unilateral market power, is the shaded area between the two marginal cost curves in Figure 2. 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.

The example given in Figure 2 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's creditors are not paid, they can demand to have the firm liquidate its assets to pay them. If a firm consistently earns revenues less than its production costs, the firm's owners and creditors will 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.

6.2. Major Dimensions of Wholesale Market Design

This section describes five essential initial conditions necessary to have competitive wholesale electricity market. Because countries have and will continue to implement wholesale markets without these initial conditions in place, I then describe a number of safeguards that limit the potential harm to consumers from implementing reforms with less-than-optimal initial conditions. I also discuss the long-term implications of these safeguards, because many of them provide short-term protection, but hinder long-term market efficiency.

As discussed previously, it is impossible to eliminate the incentive that suppliers in a competitive electricity market have to exercise unilateral market power. The best that a market designer can hope to do is reduce the incentive that a firm has to exercise this unilateral market power. 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 actions taken by the firm's competitors. As the discussion of Figure 1 demonstrates, the market designer reduces the incentive the firm has to exercise unilateral market by facing the firm with a residual demand curve that is as elastic as possible. Although I do not expect the firm's desire to maximize profits to be diminished by facing it with a more elastic residual demand curve, as Figure 1 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 financial forward contracts to supply electricity to 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 sufficient capacity to deliver electricity to all locations in the transmission network so that each firm faces 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.

6.2.1. 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 the residual demand a firm faces, 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 3 provides an example of this phenomenon. Let $SO_1(p)$ represent the bid supply curve of all other firms besides the firm under consideration and Q_d the level of demand. Figure 3(b) show 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 in at p_{max}.

To see the impact on a firm's residual demand curve from requiring it to divest capacity, suppose that the firm in Figure 3 was forced to sell off 500 MW of its capacity to a new entrant to the market. 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 3(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 3(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 is an example of a general phenomenon associated with structural divestiture, the firm 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 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 3.

6.2.2. Forward Contracts and Vesting Contracts

Much has been made of the importance of forward contracts to manage the risk of spot price volatility. However, in electricity markets forward contracts serve an even more important purpose. They make it unilateral profit-maximizing for suppliers to bid more aggressively in the spot electricity market. This point is demonstrated in detail in Wolak (2000a).

To understand the impact of forward contract commitments on supplier bidding behavior it is important to understand what a forward contract obligates a supplier to do. Usually forward contracts are signed between suppliers and load-serving entities. These contracts usually give the load-serving entity the right to buy a fixed quantity of energy at a given location at a negotiated price. Viewed from this perspective, a forward contract for supply of electricity obligates the seller to provide insurance against 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 forward financial contract to avoid any price risk is to provide the contract quantity of energy from its own generation units. This guarantees 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 (flatter bid supply functions) into the spot electricity market.

To see this point, consider the following example taken from Wolak (2000). 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 my analysis. The variable profits the firm earns during this hour are equal to

$$\pi(p) = (DR(p)-QC)(p - MC) + (PC - MC)QC. \tag{2}$$

The first term in (2) is equal to its profit or loss the firm earns from buying or selling energy in the spot market at a price of p. The second term in (2) is the variable profits the firm earns from selling QC units of energy at PC. As discussed in Section 2, 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 spot market, from the perspective of bidding into the spot market, the firm treats (PC - MC)QC as a fixed payment it will receive regardless of the spot price, p. Consequently, the firm can only impact the first term through its bidding behavior in the spot market.

Because DR(p) is downward sloping, it is possible if the market price is high, the firm will sell less energy than it forward contract commitments. However, if the price at which DR(p) is greater than QC is more than MC, the firm earns losses on the difference between QC and DR(p) times the difference between p and MC. Therefore, 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. 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 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 facing the firm and F = (PC - MC)QC, the variable profits from forward contract sales. In terms of this notation $\pi(p) = DR_C(p)(p - MC) + F$, which has exactly the same structure (except for F) as the firm's profits from selling electricity if it has no forward contract commitments. The only difference is that DR(p) replaces $DR_C(p)$ 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 expected 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:

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

where $\epsilon^c = DR_C{}'(P^c)*(P^c/DR_C(P^c))$. 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:

$$| \epsilon^c | = | DR'(p)*(p/[DR(p) - QC]) | > | \epsilon | = | DR'(p)*(p/DR(p)) |. \tag{4}$$

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 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 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 forward contract commitments between suppliers and load-serving entities, most wholesale electricity market begin operation with a large fraction of the final demand covered under 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 forward contract commitments on the part of the new owner of the generation capacity to supply a substantial fraction of the expected output of the unit to LSEs at some pre-set 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 was being sold, the regulator would assign a forward contract obligation on the new owner to supply 400 MW of energy each hour at some previously agreed upon price to one of the load-serving entities.

Vesting contracts accomplish several goals. The first is to provide price certainty for load-serving entities 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 vagaries of spot market outcomes, because they only receive or pay the spot price for production or consumption beyond the contract quantity. Finally, the existence of this forward contract obligation has beneficial impacts on the competitiveness of the spot energy market described above.

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 LSEs purchased 100% of their total energy and ancillary services requirements from the day-ahead and shorter horizon spot markets. When the amount of imports from the Pacific Northwest was substantially reduced as a result of reduced water availability during the late spring and summer of 2000, the fossil fuel suppliers found themselves facing the significantly less elastic residual demand curves for their output. This fact, documented in Wolak (2003), 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 should have and did 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.

6.2.3. Involving Final Demand

Consider an electricity market with no variation in demand or 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. This implies that there is always some likelihood that available capacity will be insufficient to meet demand.

Given available capacity, there are two ways of eliminating this imbalance, either price must be increased so as 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, the preparing for and dealing with rationing periods also causes substantial losses in economic activity.

A far 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 consumer to reduce their consumption so that supply and demand remain in balance. Although this might seem like a revolutionary concept in the electricity supply industry, this precisely how market for all other products operate.

Consumers that pay the hourly price of electricity for their consumption during the hour are not fundamentally different from generation unit owners paid according the hourly price of electricity 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." 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, as discussed above, a 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, I would expect an industrial or commercial consumer with a true supply curve of negawatts, SN(p), to bid her willingness to supply negawatts into the spot market to maximize the profits associated with selling her final output, which would imply demand-bidding to reduce the market price. Consequently, even though the generator and consumer have the sample true willingness to supply negawatts, each of them will use this true supply curve in a different manner. The supplier will use it to exercise market power on the supply side to raise market prices and the consumer will use it to exercise market power on the demand side of the market 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 4 illustrates this point. The two residual demand curves are computed for the same value of SO(p). One, Q_D, is perfectly inelastic. 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 competition 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 all 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 (2005b) 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. Some entity must manage wholesale spot price risk. Just because a state or federal regulator sets a fixed price or pattern of 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 his transmission, distribution and supply costs, must cover his 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 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 tooke over purchasing wholesale power at even higher prices. The major lesson from the California experience is that an option to purchase all of retail electricity demand at a price 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. Clearly, a more efficient way to manage electricity spot price risk is to treat consumers the same way that generation units owners are treated, because as discussed above consumers have the potential to provide the same level of grid reliability as generation unit owners.

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 they have signed a forward contract with a load-serving entity or a forward contract with some other market participant, the price they receive for any short-term energy sales is the hourly spot price. Just as very few suppliers are willing to risk selling all of their output in the spot market, I would expect consumers to have similar preferences against too much reliance on the spot 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 spot price risk that the government needs to manage using tax revenues. All consumers manage the risk of high wholesale prices, according to their preferences for taking on spot 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 spot 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 cannot purchase 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 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 such a service could be expected to be quite high. This sort of price premium should also exist for full requirements fixed-price contracts for electricity, otherwise there is an unhedged risk, that could be realized, as a number of developed and developing countries have experienced.

Borenstein (2005a) discusses a number of issues associated with involving final demand in the retail market. One roadblock to symmetric treatment of load and generation for all electricity consumers is the cost of installing the necessary metering technology at the household level to allow consumption to be measured on an hourly versus monthly basis. Wolak (2001) presents evidence for California that suggests that these costs would 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. Green and McDaniel (1998) perform a social cost-benefit analysis of the transition to retail competition for residential consumers in the England and Wales electricity market, where any consumer that wishes to switch from their default supplier must install a half-hourly meter. Green and McDaniel analyzed a number of likely scenarios for the impact of retail competition on residential consumers and electricity suppliers and found that the net benefits, if any are realized, are likely to come later as more consumers participate in the retail market and competitive pressures reduce retail prices. All of these researchers argue that there are significant benefits net of metering costs from involving commercial and industrial consumers in the wholesale market.

6.2.4. 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 a wholesale market regime. As shown in Section 6.2, in order for it to be profit maximizing for generation unit owners to submit a bid supply curve close to their marginal cost curve, they must face sufficiently elastic residual demand curves. For this to be the case, there must be enough transmission capacity into the area served by this 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 introduce the concept of 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 the local 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 regions. For example, even if the difference in the variable cost of the highest cost units operating in two regions is less than $10/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 $10/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 (lower prices) 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 for its power by 100 percent by withholding less than 10% of its capacity, it will 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 3. 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 1200 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 3 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 3 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.

6.2.5. Credible and Effective Regulatory Process

Because any attempt to establish a competitive market without the conditions outlined in previous four sub-sections is bound to result in periods when the market fails in unintended ways, it is essential that there is a credible and effective regulatory process in place to monitor the market performance to detect and correct market design flaws while they are still causing only limited consumer harm. Different from the case of the vertically-integrated utility regime, the regulator must be forward-looking and fast-acting, because, as emphasized in Section 6.1.3, markets provide extremely high-powered incentives for firm behavior, so it does not take very long for a wholesale electricity market to cause enormous consumer harm. 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 market power being exercised in the California market before it took action. In addition, the action it took was so timid and ill-conceived that its result was to increase the rate at which consumer harm occurred. Wolak, Nordhaus, and Shapiro (2000) discuss the likely impact, which also turned out to be the eventual impact, of the FERC's December of 2000 action.

An argument, based on the logic of the individual rationality constraint discussion in Section 6.1.1 can even be made that an effective, credible and fast-acting regulatory process will increase the competitiveness of a wholesale electricity market. Specifically, if the regulator makes the penalties associated with any market rule violations more than the benefits that the market participant receives from violating that market rule, then suppliers will find it unilateral profit-maximizing to obey 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 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.

FERC's failure to recognize this allowed the California electricity crisis to last as long as it did and become as big of a disaster as it did. Since the start of the California market, FERC refused to implement a system of financial penalties for market rule violations. FERC only required firms to pay back the so-called "ill-gotten" gains from market rule violations. Clearly, this an approach does not deter profitable market rule violations, because the worst case scenario for the firm is having to give the profits back and the best case is being able to keep them. Unless the regulator is flawless at detecting market rule violations, under these circumstances it is expected profit-maximizing for the firm to violate market rules because it earns zero profit if it is caught violating the rules and a positive profits when it does violate the market rules and is not caught. This is not the incentive for firm behavior a regulator wants to create. Unfortunately, this precisely the incentive that FERC created in the California market.

Any penalty mechanism the regulator implements should accomplish two goals. First, firms should pay fines for market rule violations that at least exceed the financial damages its actions impose on other market participants. Second, this penalty should also be sufficient to make the expected amount of fines the firm must pay as a result of violating a market rule greater than the expected benefit the firm obtains from this violation. This second constraint implies that the firm finds it unilateral profit-maximizing to comply with the market rules. A regulator that does not take decisive action to penalize market rule violations subject to these two constraints on the magnitude of fines imposed will soon find market rule violation more frequent, which will make it more costly for the ISO to manage the transmission network and operate its energy and ancillary services markets efficiently.

The experience of California and all other US states with wholesale markets provides another very valuable lesson for the design of an effective and credible regulatory process. Retail market regulatory policies must be consistent with wholesale market regulatory policies or wholesale market outcomes that are harmful to consumers and ultimately producers will occur. In the US this is a particularly challenging task because of the division of regulatory responsibilities between federal regulators are responsible for wholesale markets and state regulators are responsible for retail markets. The requirement to coordinate wholesale and retail market policies has a very important implication that should guide the reform process in both industrialized and developing countries. If a country does not address the five market design features described above, then it should have substantially less ambitious goals for its wholesale electricity market.

For instance, if the political process is unwilling to divest enough of the capacity of the largest supplier to new entrants, this should place limits on the form and operation of the wholesale market. If the regulator or political process is unwilling to allow retailers sufficient flexibility to manage their spot price risk or to require some or all final consumers to be treated symmetrically with generation unit owners in the wholesale market, this should constraint the type of wholesale market adopted. These constraints on the wholesale market should not be relaxed until the regulatory constraints on achieving the five goals outlined above are relaxed. Similar logic applies to a country or region that refuses to consider the economic reliability benefits of transmission upgrades in the cost-benefit calculus for transmission upgrades. A country that is unwilling to establish an independent regulator or regulatory body with the necessary statutory powers to become credible and effective should not even consider regulatory reform.

The regulatory body is the guiding force for the reform process. Unless the regulator is able to intervene and change harmful market rules or market structures, subject to judicial review, significant consumer harm is likely to occur at some point in the future.

7. Common Market Design Flaws and Their Underlying Causes

This section describes a several common market design failures and uses the framework of the previous section 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 effective local market power mitigation mechanisms, price caps and bid caps on short-term markets, capacity payment or capacity market mechanisms.

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 develop an active forward market for energy and, as discussed earlier, many of them impose minimum forward contract coverage levels of final demand at various time horizons to delivery. The spot 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.

Because the major source of benefits from electricity industry restructuring is likely to come from more efficient new generation investment decisions, rather than more efficient operation of existing generation units to meet final demand, this regulatory focus on fostering an active forward energy market is well-founded. 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 (2004) 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. These two sets of results leave open the question whether these efficiency gains were in fact passed on to consumers in the form of lower prices. Given 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, which leaves more efficient long-term investment decision-making as a major source of potential consumer benefits from restructuring.

The fundamental problem with a perspective that emphasizes short-term market design is the difficulty in establishing 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. 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 spot market power problems with 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 concentration 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.

The greater is the fraction of final demand that faces the wholesale market price, the lower is the level of final demand at which spot market power problems will arise. As discussed in the previous section, demand that pays the wholesale price is effectively a "negawatt" supplier, so the larger the share of final demand that faces hourly wholesale price, the more negawatt suppliers there are competing against the large generation unit owners attempting to raise short-term energy prices.

It is important to emphasize that having adequate generation capacity installed to serve demand according to the standards of the formervertically-integrated utility regime does very little to prevent the exercise of substantial unilateral market power in a wholesale market regime with inadequate 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 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.

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 demand that it must produce 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 that price that prevails 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 and translated into short-term energy prices.

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. This required amount of fixed-price forward contracting will not take place unless suppliers receive sufficiently highly 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.

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 purchase 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 outcome is that many suppliers had fixed-price forward contract commitments to supply energy to California LSEs, which significantly reduced their incentive to raise prices in the short-term energy market.

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 will be waiting an extremely long time. This discussion 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.

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. The typical pattern is that the political pressures make it extremely difficult to separate the former state-owned companies into a sufficiently large number of suppliers. This is followed by a period where these suppliers are able to exercise substantial unilateral market power in the short-term energy market, which leads to calls for regulatory intervention. After a period of time when these suppliers are able exercise unilateral market power, 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 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 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 for the case 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.

The experience of New Zealand is even more instructive on the issue of the need for sufficient divestiture at the start of the wholesale market. 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 was 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 when the availability of hydroelectric energy was significantly reduced relative to 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. For example, a 10 percent statewide generation ownership market share creates significantly more local market power problem when it all located in close proximity to Los Angeles or San Francisco. Consequently, regulators should recognize that one reason why suppliers might want to purchase divested units that are located in close proximity to each other is because of the greater ability to exercise local market power, although there are clearly operational cost savings associated with this geographic distribution of generation unit ownership.

Despite all of the failures associated with the divestiture process, there is one 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.[5] 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.

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 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 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 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.[6]

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 unit owners learned to predict when their unit was needed to meet a local reliability need, they very

---

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

[6]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.

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 the 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 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 a good example of an effective local market power mitigation mechanism. It applies to all units located in the PJM control 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 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 be 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) described 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 additional costs to resolve local reliability constraints would have been substantially lower and little likely not to have risen to a 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. As long as the level of the price cap is set above the variable cost of the highest cost unit necessary to meet demand, lower levels of the price 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 past two years in the California market has shown that a bid cap of $250/MWh does not impose significant reliability problems or degrade the efficiency of the spot market, because virtually all of the demand in the California ISO control area is now covered by forward contracts.

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 rather than voluntarily purchase sufficient fixed-priced forward contracts to maintain the credibility of the bid or price. The lower the price cap on the short-term market, the higher must be the mandated fraction of final demand covered by fixed-price forward contracts.

This logic implies that regulatory intervention is necessary to correct the market design flaw created by the existence of a bid or price cap on the short-term energy market. First, the ISO must have pre-specified demand curtailment plan to address the question of insufficient bids into short-term energy market (at or below the bid cap) to satisfy demand at that price. Second, the regulatory process must mandate minimum levels of fixed-price, forward contract coverage of final demand to ensure that the probability of demand curtailment is sufficiently small at the level of the price cap chosen by the political process.

7.5. Capacity Markets and Mechanisms

No topic in electricity market design more quickly separates the economists from the engineers than the issue of capacity markets and capacity payment mechanisms. Capacity market typically pay all suppliers at $/MW-month payment for capacity that can be used to serve final demand. The monthly demand for capacity for each load-serving entity is set by the ISO as some multiple of the annual, monthly or seasonable peak demand to achieve a planning reserve margin of generation capacity that serves final demand reliably. Typically, this capacity requirement is set above 110 percent and below 120 percent of a LSE's peak demand.

In the eastern US capacity markets, the capacity requirement is enforced by a penalty on retailers for failing to meet their capacity requirements. This effectively sets a price cap on the capacity market. Because the marginal cost of supplying additional capacity from an existing generation unit is zero and the demand for capacity is completely inelastic, capacity market prices tended to be extremely volatile. When no supplier is pivotal in the capacity market, the price of capacity is close to its marginal cost of zero. When one or more suppliers is pivotal in the capacity market then the price tends to equal the price cap on the capacity market. The experience of ISOs in the eastern US with capacity markets bears out this point. There is a general dissatisfaction with these capacity market among stakeholders in the eastern ISOs.

The major rationale for capacity markets in the United States appears to be a holdover from the vertically-integrated utility regime when capacity payments compensated generation units for their capital costs, because the regulatory process often used different mechanisms to reimburse unit owners for their variable operating costs. In a wholesale market regime all generation unit owners have the opportunity to earn the market-clearing price which is typically above generation unit's average variable cost when the unit is operating. In this way, the generation unit earns a return to capital during each hour it produces electricity.

This paradigm for earning a return on capital from the difference between the market price and the firm's average variable cost of production has managed to provide the appropriate incentives for investment in new productive capacity in other workably competitive industries. Moreover, the best performing wholesale electricity markets internationally in the United Kingdom, Australia and the Nordic countries do not have capacity markets or capacity payments. They have been functioning for more than ten years, longer than any market in the US, and appear to have delivered net benefits to consumers relative to the formervertically-integrated utility regime.

It is also unclear why electricity is so fundamentally different from other products that it requires paying suppliers for their generation units to exist. Consumers want cars, not automobile assembly plants. They want point-to-point air travel, not airplanes. They want a loaf of bread, not a bakery. In these markets we do not pay producers capacity payments for the fact that they own the facilities to provide the goods and services we desire. All of these industries are high fixed cost, relatively low marginal cost production processes, yet all of these firms earn their return on capital invested by selling the good that consumers want at a price above the variable cost of producing it. Clearly cars, air travel, and bread are in many way essential commodities, yet we don't need a capacity payment to ensure that there is sufficient productive capacity for these products to meet our needs.

Virtually all of the evidence from US and international markets with capacity payments has been extremely disappointing. However, this has not diminished the appetite for engineers and market operators to attempt to design new capacity markets. As noted above, it is not clear what market design problems capacity markets address. The most coherent argument given is that because it is politically impossible, particularly in the US, to run a wholesale market without a price cap, it is necessary to have a capacity market to supplement the revenues of the high variable cost peaking units. The capacity payments make up for the what is often to referred to as the "missing money' not provided by to these units by wholesale energy market sales.

Another argument is that final demand is unwilling to participate actively in the wholesale market so it is necessary to have a capacity market to reliability operate the network without an active demand side. This argument seems difficult to believe if there is no price cap on the spot market, because at some level of the short-term price, virtually all loads would not consume additional electricity and would instead prefer to sell a portion of their fixed-price forward contract commitment back to the wholesale market at a substantial profit.

Rather take a side in the debate over need for capacity payments and the design of capacity markets, I will note that, particularly in the US, this is a debate that is far from over. Two of the three eastern ISOs are currently in the midst of dramatic overalls of their capacity market mechanisms, with no end in sight for this process. Bushnell (2005) provides an insightful discussion of the rationale for capacity markets and payments and discusses the costs and benefits of the various competing models for ensuring that suppliers receive enough revenues to remain financially viable. Wolak (2004) discusses some of the shortcomings to the current capacity markets in the eastern US. Hopefully, the FERC and state public utilities commissions will learn from the experience of other countries and explore implementing aspects of the successful market designs in the UK, Australia, and the Nordic countries that do not have explicit capacity payment mechanisms.

8. Economic and Political Constraints on the Market Design Process

This section describes economic and political constraints faced by the market designer which can limit the efficiency of the wholesale market. These include the difficulty in coordinating wholesale market and retail market regulatory policies and the lack of a clear legal foundation for regulatory oversight of wholesale markets. This lack coordination between retail and wholesale market policies has implications for the design of the methodology used to upgrade the transmission network and the extent to which final demand faces the real-time wholesale price as its default price. Another limiting factor on wholesale market performance, particularly in the United States is the need for price and bids caps to limit the magnitude of spot prices. In many developing countries, the economic purchasing power of most consumers may favor a different approach to various aspects of the wholesale market design. In addition, the natural resource endowment of a country may also favor a different approach. For example, a country with abundant hydro resources may take a different approach from a country with few hydro or domestic fossil fuel resources. Finally, the legal institutions and precedents in a country may make designing an effective regulatory process for a wholesale market more or less difficult.

8.1. Failure to Coordinate Wholesale Retail Market Policies

Many of the necessary features of a successful wholesale market design require certain aspects of the retail market design to be in place. For example, a short-term wholesale market predicated on a substantial amount of final demand being covered by fixed-price forward contacts must implement the necessary retail market rules to ensure that this is in fact the case. A wholesale market design that assumes active demand side participation must have retail market policies that provide strong incentives for final demand to participate in the wholesale market. Nowhere has the disconnection between wholesale market policies and retail market policies been greater than in the US.

The former vertically integrated monopoly regime in the US electricity supply industry created limited opportunities for conflicts between state and federal regulators. As noted above, this regime involved few interstate spot market transactions of wholesale electricity, because in exchange for its legal status as monopoly for a given geographic area, each vertically-integrated monopolist had an obligation to serve all retail demand at a price set by the state PUC. States also played a major role in the transmission and generation capacity-planning decisions of its investor-owned utilities. This model for the electricity supply led to an industry dominated by state-level oversight, with most interstate electricity transactions on a long-term contract basis. The vast majority of short-term interstate electricity transactions were for reliability reasons because the state-level obligation to serve all demand at the regulated retail price imposed the entire risk of expensive spot market electricity on the investor-owned utility. In response to this state-level regulatory dominance, transmission networks were designed to serve state-level markets and interstate electricity sales for reliability reasons only.

8.1.1. Incentives and Constraints Facing FERC

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 the industry described above illustrates 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 these same entities. The merchant power producing sector has been very supportive of FERC's goal of promoting competitive 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.

8.1.2. Incentives and Constraints Facing State PUCs and Policymakers

State PUCs and policymakers face a very different set of incentives and constraints from the FERC. 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 necessary physical and regulatory infrastructure necessary for a competitive 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 rates 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. In addition, the introduction of wholesale competition also limits the scope for the state PUC and 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 competitive regime, these decision will be made by the merchant 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. The state PUCs and policymakers lose the ability to order economic transfers from certain groups of electricity producers and consumers and award them to more politically powerful or favored groups of consumers and producers of electricity. 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 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 competitive wholesale market.

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 and 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 may be very different from those the state PUC and other state policymakers would like FERC to implement.

Because FERC also decides whether wholesale rates are just and reasonable and determines what actions are appropriate to ensure that rates are just and reasonable, state PUCs and policymakers that rely only on FERC to protect consumers from the exercise of market power may be taking an unacceptably large risk. As California learned during the period June 2000 to June 2001, a long time can elapse before the necessary legal and political pressure can be brought to bear on FERC to fulfill its statutory mandate to protect consumers in a manner that state policymakers find acceptable. In the meantime, an enormous amount money can be extracted from consumers, taxpayers and the shareholders of the incumbent electricity retailers as result of the unilateral exercise of market power made possible by poorly designed state-level retail market policies.

8.2. Implications for Involving Final Demand and Transmission Planning

This regulatory conflict in the US has significantly limited the extent to which final demand has become actively involved in the wholesale market and the transmission network is planned to enhance the competitiveness of the wholesale market regime. In the name of protecting final consumers from volatile wholesale prices, in all parts of the US, all customers have the option to consume energy at a fixed retail price set by the state PUC that does not vary with hourly wholesale prices. In the name of protecting ratepayers, the methodology used by the state regulatory commissions to determine whether to undertake transmission investments do not recognize many of the competition-enhancing benefits of these upgrades.

8.2.1. Retail Market Infrastructure Essential to Support Workable Wholesale Competition

This section describes the three essential features of the state-level retail market infrastructure necessary to support a wholesale market that benefits final consumers. I first discuss the necessity of universal interval metering. Because this overhaul of the metering infrastructure will take time, I suggest a scheme for its implementation. The next subsection focuses on the role of retail competition in creating a wholesale market that benefits final electricity consumers. It describes why sustainable retail competition is impossible without the widespread implementation of interval meters. The final subsection outlines the role for state PUCs in a fully competitive retail market. PUCs must monitor the portfolio positions of all electricity retailers to ensure that they are not taking imprudent gambles in forward electricity markets that may prevent them from fulfilling their contractual obligations their retail customers.

8.2.1. The Necessity of Hourly Metering Technology

To understand the need for hourly metering technology, I first review the determinants of firm profitability in the formervertically-integrated utility regime versus the competitive wholesale market regime. I then describe why sustainable retail competition is unlikely to develop unless there is widespread implementation of hourly meters.

8.2.1.1. Firm Financial Viability Under Competition Versus Regulation

Both the regulated utility regime and the wholesale market regime require that firms obtain sufficient revenues to cover total production costs over the year from all customers. The difference between these two regimes is how firms recover these production costs.

The monopolist can serve many customers at a loss for long periods of time as long as it earns significant revenues in excess of production costs from other customers. This is possible because it has a legal monopoly on the supply of electricity for a given geographic area and other firms are prohibited by law from entering profitable segments of the monopolist's business. Consequently, the regulated regime can tolerate huge cross-subsidies from between classes of customers as distinguished by their geographic location in the transmission network, their load shape, or other observable characteristics.

Under a wholesale market regime, firms have must attempt to make profits on every KWh of energy they sell. Because there are a number of suppliers of wholesale electricity, competition among them will eventually erode any excess profit opportunities. Therefore, it is very unlikely that cross-subsidies across customer classes will persist in a competitive wholesale market.

The price a seller receives for each KWh it supplies depends on conditions in the wholesale market. Differences in competitive conditions can lead to huge variations across hours of the day, week, month or year in the price that a supplier receives for electricity sold in an hour. Different from the regulated regime, the firm may be able to take actions which impact these market conditions, and therefore the price it receives for electricity. If firm is able to this raise price more than it raises average production costs, then it is profit-maximizing for the firm to engage in activities that raise its costs of producing power.

Because firms in a wholesale market earn their profits one hour at a time, it is crucial that consumers pay for their electricity at these hourly prices. Schemes such as load-profile billing and other methods which attempt to compute proxies for a customers hourly consumption using that customer's monthly electricity consumption are largely ineffective at providing the necessary hourly price signals. The following example of load profile-billing illustrates this point.

Virtually all meters for small commercial and residential customers only can only capture usage over the time interval between 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. Under a load-profile billing scheme this monthly consumption is distributed across hours of the month according to a representative load shape proposed by the retailer and approved by the state PUC. 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 monthly consumption but shifted it during the month from hours with very high wholesale prices to those with low wholesale prices, the customer's bill would be the same.

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 competitive wholesale market the divergence between P(avg) and the actual hourly 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 thevertically-integrated utility regime, the utility received the same price for supplying electricity that the final customer paid for every KWh sold to every customer served.

The implementation of hourly meters for all classes of customers would allow prices that reflect hourly wholesale market conditions to be charged to all customers for their 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 spot 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, the major barrier being the lack of hourly meters, arbitrage should force the prices paid for both products to be equal.

One would never think of charging a generating unit anything but the real-time spot price for all energy supplied over the period that the price was valid. These prices signal the generator to when to supply more or less energy. The same logic applies to the demand side of market. But these price signals cannot operate without the ability to record the hourly consumption at the customer level.

8.2.1.2. Sustainable Retail Competition Requires Hourly meters

Hourly metering technology is crucial to the development of sustainable retail competition. 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 it below that 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.

The telecommunications industry provides an excellent example of the potential perils of introducing retail competition without the ability to measure a customer's consumption at the same level of granularity that the good is purchased in the wholesale market. Imagine running a competitive market in long-distance services with only the ability to measure the total number of minutes of phone calls a customer makes in a month. Each competitive long-distance provider would then apply a representative calling pattern set by the regulator to this monthly total number of minutes of phone calls to the customer's monthly bill.

The opportunities for customers to exploit these representative calling patterns for their own gain are enormous. A customer could claim a calling pattern with many local calls, all of short duration. However, once a representative calling pattern for monthly billing purposes was set, the customer could make long-distance calls of long duration to far away places. As long as the total number of minutes of phone calls was the same as it was under his claimed calling pattern, his monthly bill would be unchanged. However, the cost to the provider of network services in the second case is dramatically higher than it would be in the first. Clearly, no state or federal regulator would ever consider running competition in long-distance services without the ability to measure the origin and destination pair of each call, the exact time it was made and the duration of the call, all factors which determine the cost of providing this call.

What is being attempted in retail electricity competition throughout the US fits this representative calling pattern pricing model for long-distance competition. Customers are charged an average price for all of their monthly electricity consumption, regardless of when this electricity is consumed during the month. Because of the large variation in the price of wholesale power across hours of the day, week, month or year under the wholesale market regime, similar opportunities to exploit these representative load shapes exists for electricity customers. As noted above, once a customer has been assigned a representative load shape, it achieves the same monthly bill decrease from reducing its consumption by 1 KWh in any hour of the month.

In the former vertically integrated utility regime, the variation in production costs across hours of the year was significantly less than the variation in market prices in wholesale market regime. In the vertically-integrated utility regime the highest variable cost unit operating in the highest demand hour of the day was no more than three times lowest highest variable cost unit operating in lowest demand hour of the day. In the wholesale market regime, Wolak (1999) shows that the ratio of the highest to lower hourly price in the day can be order of magnitude higher. Consequently, the efficiency loss from using representative load shapes to compute a customer's monthly bill was significantly less under the vertically integrated monopoly regime relative to the wholesale market regime.

With ubiquitous hourly metering this economic efficiency loss can be eliminated by symmetric treatment of load and generation. Retail competition is an ideal mechanism for eliminating these pricing inefficiencies. 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 with the hourly pattern of electricity purchases. Rather than having to buy 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.

Universal hourly metering 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.

Conventional hourly metering technology uses mobile radio communications technology to broadcast each customer's hourly or half-hour consumption levels to a central data collection agency. This automated meter reading technology will significantly reduce the cost and time delay in the settlement process, which is the process of determining how much to charge each customer for the electricity they consume and pay each producer for electricity they provide. To those unfamiliar with the electricity supply industry, this may seem to be a relatively straightforward task. However, because of line losses throughout the transmission network generators provide more electricity to the transmission network than consumers ultimately receive from the transmission network. In addition, for a variety of reasons, meter reading errors, theft, or inefficient system operation, energy is injected into the transmission network that is never recorded as being withdrawn from the system. In the conventional monthly meter reading system, many meters are not read even on a monthly basis, which further complicates the problem of assigning the appropriate obligation to pay to each customer. As a consequence of these problems, the process of producing final settlement data against which a customer's bill is determined using conventional metering technology can take more than one year. This is the case in the England and Wales electricity market. The settlement process is not finalized until 14 months after the date the electricity was delivered.

In the former vertically integrated monopoly regime, the costs of these time lags and administrative processes was primarily borne by consumers. The monopolist received all the customer payments and the primary point of debate was whether the customer over or underpaid in a given month. However, in a competitive wholesale market where many wholesale suppliers and competitive retailers are injecting and withdrawing energy from a large number of locations in the transmission grid the magnitude of energy unaccounted for because of metering errors, load profiling errors, theft and inefficient system operation can be even greater. Moreover, unaccounted-for-energy costs impact the bottom line all competitive wholesalers and retailers, so all them can be expected to attempt to shift these costs on to other market participants. The installation of hourly metering for all customers would significantly reduce the magnitude of unaccounted for energy and costs of dealing with it.

8.2.1.3. Role of State Level Regulation of Retail Sector

The ability to trade risk among market participants is a large source of potential benefits from electricity re-structuring. In the former vertically integrated monopoly regime all of this risk was assigned to the vertically integrated utility. It managed the risk of delivering electricity to final customers in real-time and then sent the bill for doing this to consumers. In this wholesale market regime risks can be bought and sold. For that reason, we would expects risks to be borne by those entities able to manage this risk at least cost.

There is a moral hazard problem in electricity retailing similar to the one that exists in retail banking. The fear in retail banking is that the bank will take customer deposits and invest them in extremely risk assets in an effort to deliver a very favorable return to the investor and the bank's shareholders. However, engaging in this risk-taking behavior may lead to outcomes that render the bank unable to meet certain future obligations to its depositors. An analogous chain of events happens in the electricity retailing industry. The retailer has a strong incentive to under-invest in forward contracts to cover their future load obligations when it sells a fixed-price commitment to a customer for one or two-year period. It maybe able to earn a higher expected return by taking risks that increase the probability of bankruptcy but also have the prospect of very high positive profit levels due to low wholesale prices.

Consequently, similar to the retail banking sector regulation, state PUCs must monitor forward contract coverage requirements of all retailers relative to their forward retail market commitments. If firms are always required to hold a certain amount of fixed-price wholesale market commitments for given amount of fixed-price retail market commitments, then these firms will find it profit-maximizing to honor their retail market commitments.

This market monitoring process should require all retailers to submit to their state PUC on a monthly basis a list their retail market commitments by duration and price and their wholesale market coverage by quantity and price. The role of the PUC would be to verify that the retailer met these risk management prudency standards and assess penalties and sanctions for violations.

Consider the following example of how this might work. The second and third column of Table 1 contains a list of the quantity-weighted average wholesale price implicit in the fixed retail price retail and quantity obligations that the retailer has agreed to supply for various delivery months in the future. The fourth and fifth columns of Table 1 contain the quantity-weighted average fixed wholesale price and quantity commitments the retailer has signed with wholesale energy suppliers. The sixth columns contain the desired percentage of the total monthly quantity of fixed-price wholesale quantity commitments that the state PUC deems that it prudent for the retailer to hold as a hedge against its fixed price retail commitments for each future delivery date. The last column contains the product of the percentage in the sixth column and the fixed price retail obligation quantity given in the second column.

In this example there are several horizons where the desired hedge quantity is greater than the amount given in the fourth column. In these instances there are several actions that the state could take. First, it could assess a penalty per MWH on the positive part of difference between desired quantity in the seventh column and the actual quantity in the fourth column. The PUC could also prohibit this retailer from selling more fixed-price retail obligations at this time horizon or shorter until the retailer submits a monthly report that is not out of violation for all months longer than this delivery horizon.

For the case given in Table 1, the first month the retailer is out of compliance is month 4. This means that retailer is prohibited from signing fixed price commitments for deliveries longer than 3 months in the future during the next month unless it submits proof of compliance in the next month for all delivery horizons up to 3 months. There are other prudency standards that state PUCs could impose on hedging behavior of retailers that uses risk measures based on the prices of retail obligation versus the price of wholesale commitments that cover them. Fortunately, these hedging standards do not need to be set using very sophisticated methods in order provide a reasonable level of assurance that all retailers will be able to meet their fixed price retail obligations with a high degree of certainty.

The other role of the state PUC in a competitive retail market is to ensure that all retailers have equal access to the billing and metering services provided by the regulated monopoly local distribution company. The PUC must establish rules that prevent the local distribution company from favoring its competitive retailing affiliate.

The technologies of hourly meter installation, operation, and maintenance imply that the most cost effective strategy for quickly creating a market with ubiquitous hourly meters is to make the provision of metering and data collection services part of the local distribution company's regulated services. The regulated distribution company would formulate jointly with the state PUC an aggressive plan for installation of hourly metering for all customers in the most cost-effective manner possible. At the same time the local distribution company can manage the elimination of meter reading and other jobs associated with manual meter reading and data processing.

8.3.1. Competition Enhancing Transmission Expansion Policies

In the former vertically integrated regime, the utility would undertake transmission expansions if this could reduce the cost of serving its customers. In the wholesale market regime, an expansion in one portion of the network could have limited or no benefits to the customers located near the expansion, but have large benefits to the customers distant from the expansion. Because they were designed serve the former vertically-integrated regime, state-level regulatory processes for assessing the benefits of transmission expansions typically ignore these benefits in assessing the financial viability of transmission upgrades. Because of the political pressures faced by state regulators described above, they find it very difficult to consider benefits accruing outside of their jurisdiction in assessing the viability of a proposed upgrade.

Another important rationale for transmission upgrades during the transition to a formal wholesale electricity market is to ensure that all consumers, regardless of their location in the transmission network, receive a share of the benefits from this new industry structure. Specifically, operating a wholesale electricity market using a transmission network built by the former vertically-integrated utility can impose enormous economic burden on certain groups of consumers through no fault of their own, while at the same time allowing other consumers to realize significant benefits.

The former vertically-integrated utility had a strong incentive to build a transmission network and locate generation units to minimize the total cost of meeting all demand in its service territory. This incentive implies that the utility might install a small high-cost generation unit in an area rather than upgrade the transmission network into that region because this was the least cost way for the utility to meet an annual increase in demand in an isolated portion of its service territory.

However, in a wholesale market regime, where each consumer pays the wholesale price of energy at their location in the transmission network, consumers located in this isolated geographic area will pay a price for all of their consumption set by this high cost unit. In the former vertically-integrated regime, the cost of operating this unit was averaged over all of the electricity sold by the vertically-integrated utility to set the price and all consumers were charged for electricity.

Although all consumers should pay the market price of energy at their location in the transmission network, it important to emphasize that the reason some consumers pay much higher prices in a wholesale market regime is because the existing transmission network was designed to serve the vertically-integrated utility regime. The network was not build to facilitate a competitive wholesale market at all locations in the transmission network, but instead to minimize the overall costs of meeting the vertically-integrated utility's retail load obligations.

Consequently, another argument revising the methodologies used by state PUCs to assess the viability of transmission upgrades is to limit the magnitude of inequities in the allocation of system-wide wholesale energy purchase costs to consumers due to the past transmission and generation investment decisions by the vertically-integrated utility. Those consumers facing extremely high local energy costs because of the transmission and generation investment decisions of the former vertically-integrated utility should receive transmission upgrades to increase their access lower cost distant suppliers, and therefore share in the benefits of wholesale electricity competition.

Unfortunately, electricity industry restructuring has effectively severed the incentive to undertake transmission upgrades from the ability to do so. Generation-unit owners profiting from congestion, as described earlier, have no incentive to support the upgrade. Electricity retailers bundle these congestion charges into their cost of purchasing wholesale electricity. Transmission owners receive a regulated rate of return on their network investments. Only consumers would like economically beneficial upgrades to occur, but individually they have little incentive to participate in the process. If anything there appears to be a greater need than in the former regime for coodination of state and federal oversight of transmission expansions in the wholesale market regime to realize benefits of wholesale competition.

9. Explaining the US Experience

This section uses the results of the previous three 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 will be 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.

9.1. 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 pricing plans.

The lesson from California is that once a state introduces a wholesale market with a significant merchant generation segment–generation owners will 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.[7] Typically the vertically-integrated utilities retained a substantial amount, if not all of their generation capacity in a 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 a long-duration fixed-price forward contracts. As a consequence of these decisions, the eastern ISOs had very few generation owners with substantial net long position relative their retail load obligations. Consequently, these market were less susceptible to the exercise of unilateral market power at intermediate load conditions, relative to California, where virtually all of the output of the merchant sector was purchased from California's short-term energy markets.

Forming the eastern ISOs from tight power pools retained many vestiges of the former vertically-integrated utility regime, which in many cases increased costs to consumers from restructuring. In particular, all of the eastern pools power have capacity payments mechanisms which make much more sense in a vertically-integrated regulated regime with a power pool power. A capacity payment mechanism is useful to recover going forward fixed costs and a return on capital invested. Transferring these same capacity payment mechanisms to a wholesale market regime where all suppliers are paid the market-clearing price at their location for all energy they inject into the network can significantly increase the overall cost to consumers, unless the operating efficiency

[7] 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.

gains from restructuring are sufficiently large. This difference in revenue recovery from simply paying the total variable cost for the energy produced and paying the same market-clearing price to all energy produced can be significant, particularly in regions with relatively steep system-wide marginal cost curves. The regulatory overhang from the former pool power regime in a some the eastern markets was substantial. For example, at the start of the New England ISO, there were two capacity markets–an installed capacity markets (ICAP) and an operable capacity market (OCAP). The complications and expense of operating these two markets quickly caused the New England ISO to eliminate the OCAP market.

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

Another reason for the relative poor experience of the US versus 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 significant 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 formervertically-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 impressive 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.

9.3. 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 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 an individual generation unit bids above its incremental cost, other more expensive units may be chosen to supply in its 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, assuming that a scarcity pricing mechanism is in place that would raise prices enough for demand to clear the market.

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.

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 they should receive cost recovery. But cost recovery need not distort prices in periods or at locations where there is no justification for prices 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. The 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. The 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 gets prices wrong all the time, producing the problems described above.

9.4. 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 better suited to a wholesale market regime than that transmission networks in the US.

9.5. 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.

One way this manifests itself at FERC is a concern that all generation investment must obtain full cost recovery. This logic makes more sense in the formervertically-integrated utility regime where all investment decisions by the utility were approved by the state PUC. A discussed earlier, this significantly increases the likelihood that investments that turn out to be ex post imprudent are still paid for by ratepayers.

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 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 charges 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

The regulatory process necessary for a successful restructuring process in the electricity supply industry may be practically impossible to achieve in the US. Much more so that in the former vertically-integrated utility regime, wholesale and retail market policies must be extremely well-matched in a restructure 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 so by how FERC is generally perceived by state PUCs to have handled the California electricity crisis. These relationships appear to have improved over the two years as a result of number of positive changes at FERC, but there is 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.

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 (2005a) “Time-Varying Retail Electricity Prices: Theory and Practice,” in Electricity Deregulation: Choices and Challenges (edited by) James Griffin and Steven Puller, University of Chicago Press: Chicago, IL.

Borenstein, Severin (2005b) “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

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 (2004) “Does Competition Reduce Costs? Assessing the Impact of Regulatory Restructuring on U.S. Electric Generation Efficiency,” December, Center for the Study of Energy Markets Working Paper Number CSEMWP-153R, available from http://www.ucei.berkeley.edu/pubs-csemwp.html

Green, Richard and McDaniel, Tanga (1998) "Competition in Electricity Supply: will "1998" be worth it?" April, POWER working paper, PWP-057, available from http://www.ucei.berkeley.edu/PDF/pwp057.pdf

Green, R.J. (1999) "The Electricity Contract Market in England and Wales" Journal of Industrial Economics, Vol XLVII, No 1, pp.107-124

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 (2000) “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 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.

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.

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) “What’s Wrong with Capacity Markets?” June. (*)

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.

Table 1: Sample Monthly Forward Contract Levels Compliance Filing						
	Retail Obligations		Wholesale Purchases		Compliance Levels	
Future Delivery Date for Energy (months)	Total Quantity (MWH)	Average Implicit Wholesale Price ($/MWH)	Total Quantity (MWH)	Average Purchase Price ($/MWH)	Hedge Factor (%)	Desired Hedge Quantity (MWH)
1	10000	44.56	9400	40.12	90	9000
2	10000	45.60	9400	45.00	90	9000
3	10000	42.00	9600	40.21	90	9000
4	12000	50.00	9800	49.00	85	10200
5	13000	54.00	8600	52.00	85	11050
6	11000	51.00	8000	50.12	80	8800
12	10000	48.00	8000	45.29	80	8000
18	10000	44.23	8000	39.56	75	7500
24	12000	44.00	8000	42.03	70	8400

Figure 1: Residual Demand Elasticity and Profit-Maximizing Behavior

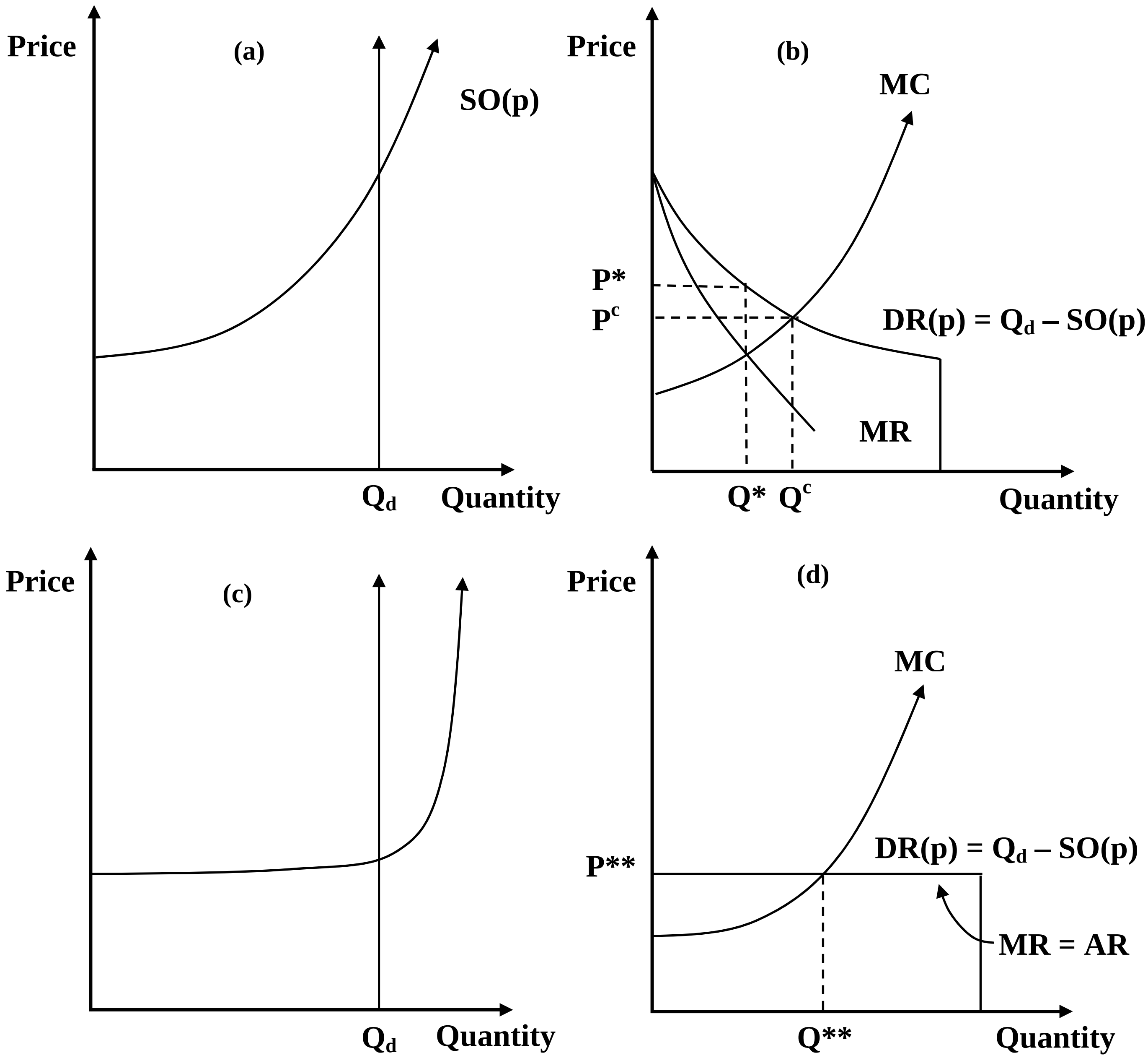

Figure 2: Welfare Loss from Inefficient Production

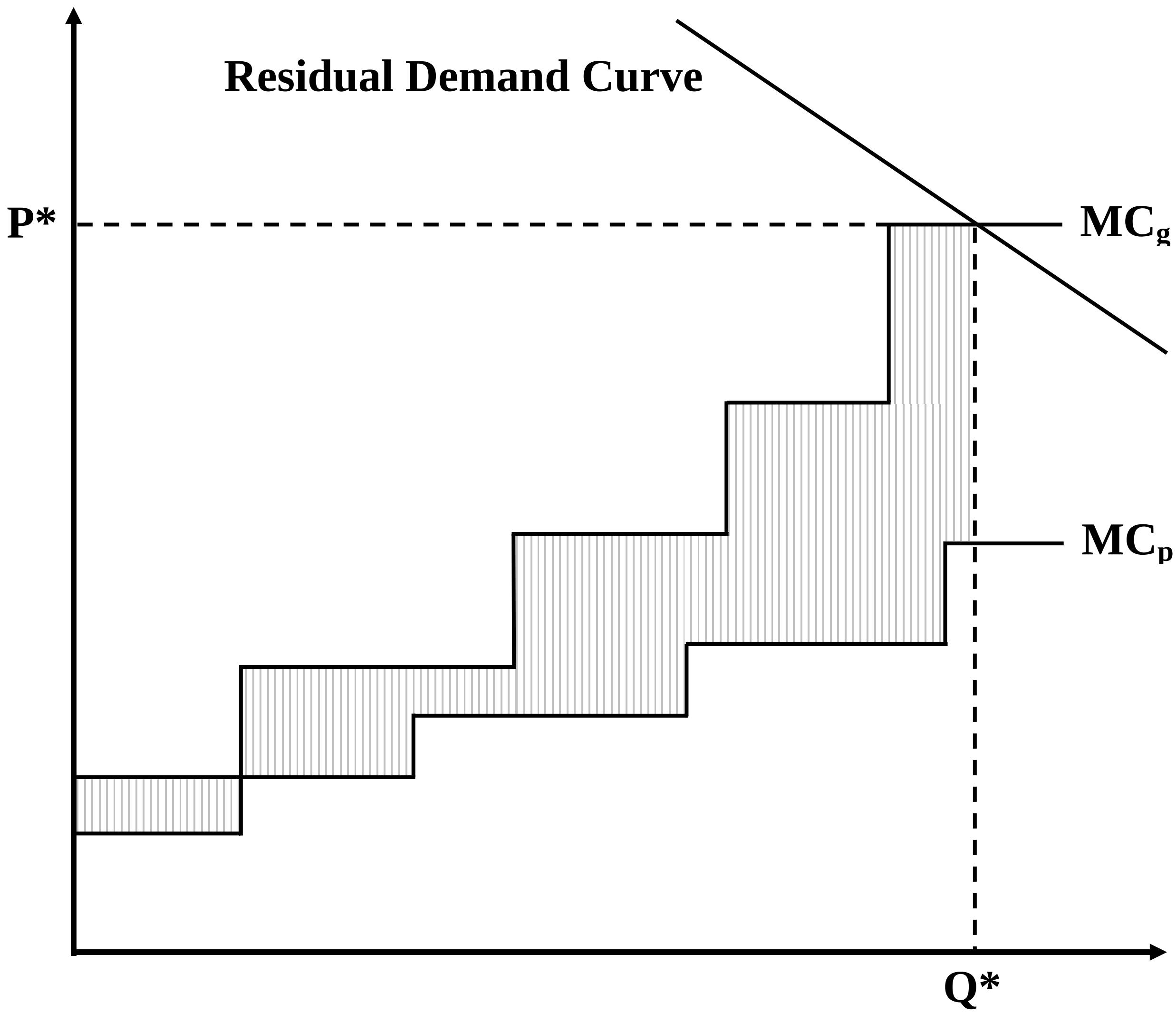

Figure 3: The Impact of Capacity Divestiture on a Pivotal Supplier

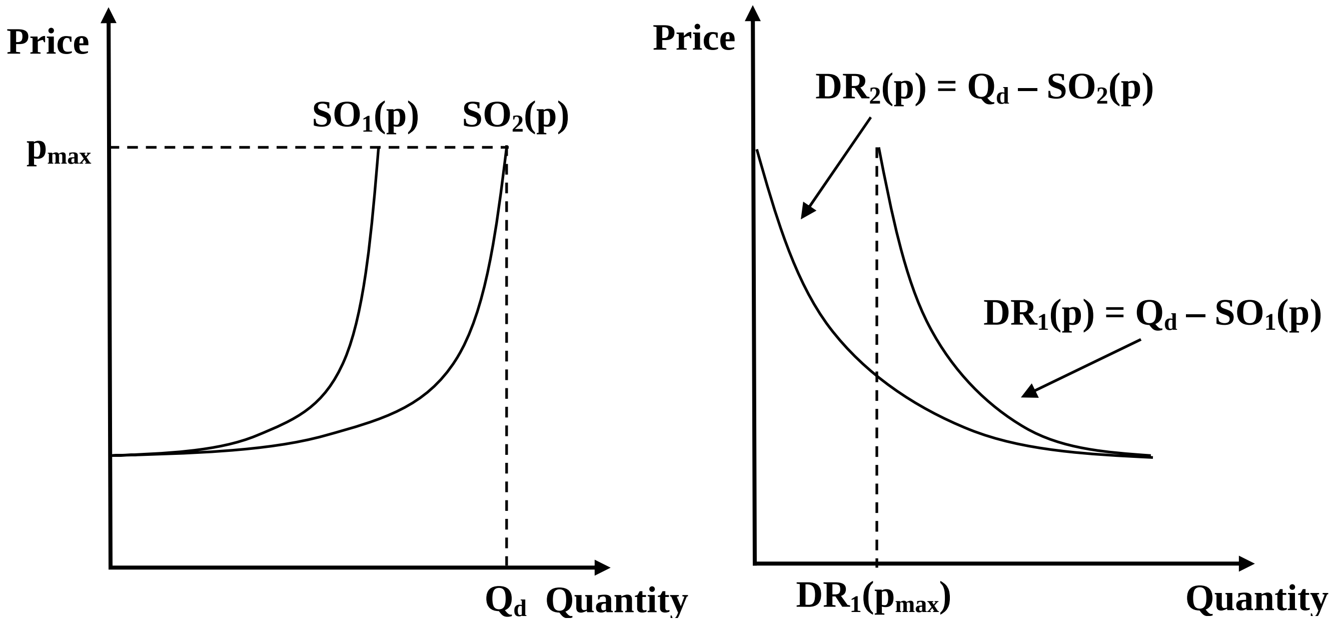

Figure 4: Residual Demand Elasticity and Price-Responsive Demand