December 5, 2007 (Revised)

CAPACITY PAYMENTS IN IMPERFECT ELECTRICITY MARKETS: NEED AND DESIGN[1]

Paul L. Joskow
Professor of Economics and Management
Department of Economics
Massachusetts Institute of Technology
50 Memorial Drive, E52-271D
Cambridge, MA 02142
USA

ABSTRACT

This paper argues that a variety of imperfections in wholesale “energy-only” electricity markets lead to generators earning net revenues that are inadequate to support investment in a least cost portfolio of generating capacity and to satisfy consumer preferences for reliability. Theoretical and numerical examples are used to illustrate the sources of this “missing money” problem. Improvements in “energy-only” wholesale electricity markets, especially those that improve pricing when capacity is fully utilized, can reduce the magnitude of the missing money problem. However, these improvements are unlikely to fully ameliorate it. Forward capacity obligations and associated auction mechanisms to determine capacity prices are necessary to restore appropriate wholesale market prices and associated investment incentives to support the optimal portfolio of generating investments. The deficiencies of the original capacity payment mechanisms adopted in the U.S. are discussed and the necessary improvements identified.

1. INTRODUCTION

A lot has been learned about the effective design of wholesale electricity market institutions over the last two decades. I believe that we now understand the key elements of wholesale electricity market designs that work reasonably well to induce efficient dispatch of generators, efficient management of network congestion, that provide incentives for generators to operate efficiently, that yield efficient prices for generation services, and that supports reliable network operations. That is, we now understand how to design wholesale electricity markets that work well in the short run. The primary

[1] An expanded discussion of the issues addressed in this paper can be found in Joskow (2007) and Joskow and Tirole (2006, 2007) upon which this paper relies.

barriers to effective wholesale market design to support competitive wholesale markets with good short run performance attributes are political rather than intellectual.

1.1 Wholesale Markets Design for Good Short Run Performance

The basic wholesale market design features that lead to good short term performance include: the creation of *voluntary* transparent organized spot markets for energy and ancillary services (day-ahead and real-time balancing) that accommodate bilateral contracts and self-scheduling of generation if suppliers choose to take this approach; locational pricing of energy and ancillary services reflecting the marginal (opportunity) cost of congestion and losses at each location; the integration of spot wholesale markets for energy with those for ancillary network support services and with the efficient day ahead and real-time allocation of scarce transmission capacity; auctioning of financial transmission rights that are simultaneously feasible under alternative system conditions to hedge congestion, serve as a basis for incentives for good performance by system operators and transmission owners, and partially to support new transmission investment[2] (Joskow and Tirole 2000, Gilbert, Neuhoff, and Newbery 2004); an active demand side that can respond to spot market price signals (Borenstein, Jaske and Rosenfeld 2002) and can be used by system operators to manage network reliability. These are the attributes of the wholesale electricity markets that are now operating in the PJM, New England, New York and the Midwest ISO areas of the U.S. (Joskow 2006). California is proposing to implement a similar “nodal price” market design, though its implementation has been delayed until 2008. Texas is considering adopting important features of this wholesale market design as well. I will refer to

[2] The allocation of transmission rights can, however, affect the incentives of firms to exercise market power and this should be taken into account in the design of rights allocation mechanisms and restrictions on the entities that can purchase these rights (Joskow and Tirole (2000), Gilbert, Neuhoff and Newbery (2002)).

market designs that have these attributes as the "standard 'energy only' market design" in what follows.

Wholesale markets that have adopted and refined the standard energy only market design work quite well under most contingencies (Joskow 2006a). While markets without transparent locational pricing and which do not integrate spot energy markets and ancillary services markets with network congestion management in a transparent fashion can work reasonably well (e.g. NETA in the UK), there are costs associated with not adopting locational pricing (Green 2004). Moreover, markets that have not adopted the primary standard market design features are more likely to run into serious operating problems and to lead to disputes between stakeholders (as in California, Texas, Alberta and Ontario) arising from the administrative non-price allocation of scarce transmission capacity and the pricing of generation service when the transmission network is constrained (see the chapters on these markets in Sioshansi and Pfaffenberger 2006).

1.2 Wholesale Market Design: Scarcity Pricing and Investment Incentives

There is one set of contingencies where the implementation of the standard "energy only" market design does not yet work well. These are conditions when demand is at or near its peak level and generating capacity is fully utilized. By "fully utilized" I mean that all of the capacity available on the system is needed to supply energy or ancillary services to balance supply and demand consistent with reliability criteria imposed on the system operator. Ideally, under these "scarcity" conditions prices for energy and ancillary services would rise to clear the market consistent with maintaining network reliability. Specifically, wholesale prices would rise to reflect the opportunity cost of a network failure or the value of lost load VOLL (Stoft 2002, Joskow and Tirole

2007). In practice, in most of these wholesale markets, wholesale prices for energy do not rise fast enough or high enough to clear the market and maintain network reliability. System operators rely instead on actions that increase the probability of a network collapse, non-price rationing of demand, and out of market (OOM) bilateral arrangements with certain generators to balance supply and demand.

There are several reasons for this (see Joskow 2006 and 2007 for a more detailed discussion): (a) the quantity of day-ahead and real time demand response available to the system operator to clear the market and maintain network reliability is too small; (b) to maintain network reliability under these conditions system operators reduce the level of system security by using operating reserves to supply energy, increasing the probability of a network collapse (Joskow and 2007), and the associated social costs of increasing the probability of a network collapse are difficult to get reflected in market prices; (c) and/or system operators must (or think they must) resort to non-price rationing (rolling blackouts) or other "out of market" actions to balance the system in anticipation of problems later in the day and before prices rise to reflect the anticipated scarcity; (d) regulators impose administrative price caps placed on prices for energy and ancillary services to deal with potential market power problems that are far below the VOLL that would clear the market when capacity is fully utilized ; and (e) other emergency reliability protocols used by system operators affect (suppress) market prices in ways that create a wedge between wholesale market prices and the social costs of the system operator's "out of market" actions. Indeed, efficient prices that reflect the social costs of some emergency actions (e.g. voltage reductions) may not be able to be implemented by a market mechanism.

1.3 Scarcity Conditions and Scarcity Pricing

There are only a small number of “super-peak” demand hours (e.g. 10 to 50) each year when capacity is fully utilized and operating reserve deficiency and other reliability protocols must be implemented on the typical system. As long as system operators use reasonably sensible non-price rationing schemes, and give advanced notice of rolling blackouts wherever possible, the short run costs of these imperfections in wholesale market institutions are likely to be fairly small, unless there is a wide scale system collapse, something that is extremely rare in developed countries. However, these market imperfections create long run inefficiencies as well, and these inefficiencies are likely to be much larger. These long run inefficiencies are associated with the failure of wholesale market prices for energy and ancillary services to rise high enough to clear the market when capacity is fully utilized to induce efficient levels of investment in new generating capacity, consistent with the costs of different types of generating capacity and consumer valuations for reliability..

Because electricity cannot be stored economically and electricity demand varies widely over the hours of the year, sufficient capacity must be built to balance supply and demand reliably under peak demand conditions. This implies as well that a significant amount of generating capacity on an efficient system is “in the money” to generate electricity for only a small fraction of the hours during a typical year. The last increment of generating capacity may not be called at all to generator electricity in many years, standing in reserve to meet low-probability high-demand contingencies. This means that these generators must earn all of the net revenues (revenues net of fuel and other operating costs) required to cover their investment costs during these few critical hours.

To do so, energy and ancillary service prices must be quite high during these hours in an "energy only" market design (i.e. without capacity payments) to induce investment in generation consistent with the reliability criteria imposed on system operators. Infra-marginal generators in an efficient generation portfolio may earn a significant fraction of their net revenues during these hours as well. If prices during these few critical hours are too low, then the net revenues will be inadequate to support the efficient quantity and mix of generating capacity; that is, there will be underinvestment in generating capacity, too many hours when capacity is fully utilized, too much reliance on non-price rationing, and too high a probability of a network collapse. I will follow Cramton and Stoft (2006) and refer to this as the "missing money" problem.

There is substantial empirical evidence that the markets in the U.S. that have adopted the standard market design suffer from the missing money problem (Joskow 2006, 2007). The "missing money" problem will be the focus of the rest of this paper.

1.4 Investment Incentives: Additional Considerations

Before proceeding, however, let me note that there are other arguments that have been advanced for why competitive wholesale electricity markets will necessarily lead to underinvestment in generating capacity and/or to an inefficient mix of generating capacity. It is sometimes argued that short-term wholesale electricity prices are too volatile to support new investment in long-lived capital intensive generating capacity without support from long term contractual agreements between generators and wholesale or retail supply intermediaries. Retail customers, with a few exceptions, show little interest in entering into contracts of more than two or three years duration and, for this and perhaps other reasons, a liquid voluntary forward market for longer duration contracts that investors can rely on to hedge electricity market risks has not emerged naturally. A variant of this “uncertainty barrier” argument is that the problem is not that investments will not be forthcoming at some price level, but rather that the cost of capital used by investors to evaluate investments in new generating capacity that will operate in competitive wholesale spot markets for energy and operating reserves is so high that it implies electricity prices that are even higher than those that would have been experienced under the old regime of regulated vertically integrated utilities where market, construction, and generator performance risks are largely shifted to consumers by fiat through the regulatory process. This then turns into an argument against liberalized electricity sectors.

Second, it is also sometimes argued that market rules and market institutions change so frequently and that opportunities for regulators to “hold-up” incumbents by imposing new market or regulatory constraints on market prices is so great that uncertainty about future government policies acts as a deterrent to new investment. This is especially problematic in electricity markets because a large fraction of the net revenues earned to compensate investors for the capital they have committed to generating capacity relies on very high spot market prices realized during a very small number of hours each year. The potential opportunity for market rules and regulatory actions to keep prices from rising to their appropriate levels even in a few hours each year when efficient prices would be very high can seriously undermine investment incentives.

The first argument has little merit. Large investments in production facilities whose output exhibits significant price volatility occur all the time (e.g. oil and natural gas). The second argument has more merit, as policymakers have not been shy about ex post adjustments in electricity market designs and residual regulatory mechanisms, sometimes motivated by a desire to hold up existing generators opportunistically. I have discussed these arguments in more detail elsewhere (Joskow 2007) and focus on the missing money problem here.

2. A BENCHMARK MODEL OF AN "IDEAL" WHOLESALE ELECTRICITY MARKET

It is useful to start by characterizing a simple model of an "ideal" wholesale electricity market that does not suffer from the problems noted above and, in particular, where there is no "missing money" problem. Joskow and Tirole (2007) specify such a simple theoretical model of a wholesale electricity market with the following attributes:

a. There is a continuum of demand contingencies with a known probability distribution;

b. There are both price-sensitive consumers (on real time meters) and price insensitive consumers (on traditional meters) served by retail suppliers (LSEs).

c. Retail suppliers can offer consumers contracts that specify the conditions under which they will be rationed.

d. The technology for generating electricity is characterized by investment costs per unit of output defined by I(c) and marginal operating costs defined by c. I(c) is strictly decreasing in c (so unit investment cost declines as marginal operating cost increases).[3] The most interesting case is where the generation technology has an upper

[3] In the model uncertainty is introduced on the demand side, but it can also be added to the supply side without loss of generality by introducing generating unit outage probabilities or availability factors. Generating unit outages are considered later in Joskow and Tirole (2007) when they add operating reserves to the model.

and lower bound on I(c) and an associated lower and upper bound on marginal generating cost c.

Under these assumptions, Joskow and Tirole (2007) show that the second-best (given price-insensitive customers and non-price rationing) optimum is given by the following conditions:

a. Wholesale prices that reflect the marginal cost of generation except when there is rationing in which case the price jumps to VOLL.[4]

b. There is efficient dispatching of generation. Generating capacity is fully utilized from lowest marginal operating costs up to the point where $p_i < c$.

c. Price sensitive consumers on real time meters are never rationed, responding to real time prices based on their preferences.

d. Price insensitive consumers on traditional meters may be rationed and the magnitude of the efficiency loss from non-price rationing depends on when and how they are rationed by their LSE.

e. Efficient investment is characterized by investment up to the point where $I(c) = (p_i – c)$, where $(p_i – c)$ defines the quasi rents earned by the marginal increment of generating capacity under demand contingency i.

It is this last condition that is most relevant for the issues of concern here. If wholesale prices are distorted, for example by placing a cap on prices so that they cannot clear the market efficiently under all contingencies, then the quasi-rents available to cover the costs of investment in generating capacity will be distorted and, in turn, investment in generation will be distorted as well. In the typical case with binding price

[4] When operating reserves are introduced the price rises monotonically from a level equal to the marginal cost of the last unit (highest c) dispatched on the system to the VOLL as operating reserves are depleted.

caps that constrain prices[5] from rising to their competitive levels under peak demand contingencies, the effect is to keep prices too low, yielding underinvestment in generating capacity and excessive rationing of consumers;[6] (p_i – c) will be too small under some contingencies and this is the source of the "missing money" problem.

3. THE BENCHMARK MODEL: NUMERICAL EXAMPLES[7]

The simple economics of the efficient utilization, investment and pricing for an electric generating system developed by Joskow and Tirole (2007) is usefully clarified with a few simple numerical examples presented in Joskow (2007). These examples ignore uncertainty on the demand and supply sides for simplicity.

Table 1 displays the parameters of three hypothetical electric generating technologies with different capital cost/operating cost ratios. The capital costs of a generating facility are fixed costs once the investment to build it has been made. The operating costs vary directly with the production of electrical energy from the generating facility.[8] There is a "base load" technology with relatively high capital costs (annualized) and low operating costs. Next there is an "intermediate load" technology with lower capital costs and higher operating costs. Third, there is a "peaking" technology with still lower capital costs and higher operating costs.

[5] If the price caps are so cleverly designed that they only constrain prices that reflect market power and being them to competitive levels then there is no distortion. However, since price caps in the U.S. are set far below estimates of VOLL, it must be the case that under some contingencies prices are not allowed to reach their efficient competitive levels.

[6] If prices are constrained in this way it is likely to be efficient to ration both price sensitive and price-insensitive consumers.

[7] These examples and the associated discussion of investment and dispatch behavior should be familiar to anyone who has read the old literature on peak load pricing and investment for electricity. See for example, Turvey (1968), Boiteux (1951, 1960), Joskow (1976), Crew and Kleinfdorfer (1976). Well functioning markets should reproduced these idealized "central planning" results.

[8] For the purposes of this example we will ignore so-called fixed operation and maintenance expenses which are incurred each year simply to keep the plant available to produce electricity after the initial investment in it has been sunk.

In order, to introduce demand-side response in a simple way into the example, it is convenient to conceptualize “demand response” as a “generation” technology option through which demand is paid to reduce consumption. This turns out as well to be the way that system operators often think about demand response under operational conditions. The payments to induce demand response are set to reflect the marginal value consumers place on consuming less energy in the very short run or VOLL (See Stoft (2002), Chapter 2-5). The numerical example in Table 1 uses a constant value of $4000/MWh for VOLL.[9] This value is well within the range of available estimates used in practical applications (e.g. in the old E&W pool and in Australia) and estimated in the literature (Joskow 2007).

Finally, a conventional inelastic load duration curve is defined to close the example. (Demand elasticity has already been introduced with the demand response technology.) In the example, demand is less than or equal to 10,000 MW for the entire year (8760 hours) and is 22,000 MW for only one instant during the year. System demands between 10,000 and 22,000 MW are realized for between 8760 and one instant during the year.

3.1 Least-Cost Portfolio of Generating Capacity

Since electricity cannot be stored economically and markets must clear continuously, total costs (capital plus operating) per unit of generating capacity for each technology varies with the number of hours that the capacity is utilized to produce electricity each year. More importantly, from an investor's perspective the comparative total costs of the three technologies depends in part upon how many hours each year it is

[9] I assume VOLL is constant for expositional and computational simplicity. Conceptually VOLL represents a demand function that would have different values reflecting diverse consumer values for electricity and an associated aggregate demand elasticity.

anticipated that each will be economical to "dispatch" to supply electricity. If a generating unit is expected to operate economically (profitably) for 8760 hours per year, the base load technology in the example is the lowest cost choice. If a generating unit is expected to be economical (profitable) to run, for example, only 4,000 hours per year, then intermediate load technology in the example is the lowest cost choice. If the capacity is expected to be economical (profitable) to run, for example, 200 hours per year, then the peaking technology is the least cost option.

For this example, Table 2 displays the least costs portfolio of generating capacity and demand response, the total costs (operating plus capital) for each technology and for the system in the aggregate, and the most efficient utilization duration (running hours) for each technology consistent with the generating technology, demand response, and load duration parameters in Table 1. In this example, the least cost portfolio includes a lot of base load capacity, a much smaller amount of intermediate capacity, an even smaller amount of peaking capacity, and a very small number of hours when conventional generation is fully utilized and "demand response" is called upon and paid to clear the market.

3.2 Implementation of the Least-Cost Generating Portfolio with Competitive Markets

One can think of the generating investment and utilization program displayed in Table 2 as what a well-informed benevolent social planner would come up with. That is, this is a benchmark result against which wholesale electricity market behavior and performance can be compared and is a simple example of the optimality conditions for the model developed in Joskow and Tirole (2007). The question then for evaluating the behavior and performance of a competitive wholesale electricity market is whether and how market prices can provide incentives for decentralized decisions by profit-maximizing investors to replicate the efficient investment and utilization program.

Table 3 displays the number of hours that each technology is the marginal supplier and the associated competitive market prices. It should be obviously immediately that except when demand reaches 22,000 MW and fully utilizes all of the generating capacity in the least cost program, that the electricity market will operate in a regime where there is “excess” conventional generating capacity. When demand is less than or equal to 14,694 MW, base load capacity is marginal and the perfectly competitive market price will be $20/MWh. When demand lies between 14,694 MW and 19,511 MW the marginal unit is the intermediate technology and the perfectly competitive market price will be $35/MWh. Finally, as demand rises above 19,511 MW and until is reaches 21,972 MW, peaking capacity is marginal and the perfectly competitive market price will be $80/MWh. Beyond 21,972 MW it is economical to call on the demand response technology by allowing the market price to rise to $4000/MWh. That is, when “price-insensitive” demand would otherwise rise above 21,972 MW, the efficient price jumps from $80 to $4000/MWh to induce efficient demand response. This is more efficient than building more peaking capacity to balance supply and demand.

Table 4 displays the revenues, total costs and any differences between revenues and total costs (shortfall or net revenue gap) for each technology and in the aggregate if the efficient wholesale market prices are realized. It shows that all of the costs of the efficient investment program covers are covered from wholesale market revenues, both for each technology and in the aggregate if efficient wholesale market prices are established. That is, competitive pricing satisfies a break-even constraint for the least-cost investment program. The efficient investment program can therefore be implemented by a competitive wholesale market that exhibits the efficient pattern of prices.

Table 5 shows just how important is “scarcity pricing” that induces demand response for achieving the break-even constraint for the optimal program. For base load technologies 33% of the quasi-rents produced to cover capital costs come from scarcity pricing, for intermediate load technology 50% of the quasi-rents come from the few hours of scarcity pricing, and from peaking technology 100%. Without scarcity pricing to induce demand response the least cost investment program would not be profitable and could not be implemented by a competitive market.

3.3 Breakeven Conditions Absent Scarcity Pricing

Assume that a regulator, having taken an elementary economics course in college and recalling the simple competitive market rule that “price should equal marginal cost,” determines that wholesale prices for electricity should never rise about $80/MWh, the marginal generation cost of the generating capacity at the top of the generation supply stack. She concludes that higher prices must reflect market power and sets a price cap at $80/MWh. That is, scarcity pricing is not permitted to occur. Table 6 shows the impact of such a policy on the profitability of the optimal generation investment program. It should be clear that short run marginal cost pricing yields revenues that are not nearly adequate to cover the total costs for any technology or total generating costs in the aggregate at the efficient investment levels. The shortfall turns out to be $80,000/MW of installed capacity for all technologies. Note for future reference, that the $80,000/MW of generating capacity required to meet a breakeven constraint for the efficient investment program is also exactly equal to the annualized capital charges for a MW of peaking capacity. Clearly, decentralized markets will not attract investment to support a least cost generation investment portfolio under this short run marginal cost pricing scenario since it would be unprofitable. For investors to break even, the market must somehow come up with another $80,000/MW of generating capacity, or $1.760 billion (an increase in total revenue of 30%). This is an extreme example of the missing money problem.

The failure to include active price-related demand response in this way or by keeping prices from rising to $4000 under scarcity conditions through some other mechanism, for example by imposing price caps that are greater than $80/MWh but less than $4000/MWh, does not imply that no investment will be profitable. Rather it implies that the efficient quantity and mix of generating capacity will not be profitable and, in a market context, an efficient investment program would not be sustainable. Less investment will be forthcoming and their will be more hours when the system operator must deal with an excess demand situation. Absent price-related demand response, the system operator will have to find some alternative way to ration demand at the time of system peak and, since both supply and demand will be vertical under this contingency, define some default price at which suppliers will be compensated for energy and operating reserves since under these conditions a market clearing price is indeterminate without demand response.

3.4 Scarcity Pricing, Investment and Reliability

Free entry into the competitive wholesale market implies that investment in generating capacity will adapt to whatever default pricing arrangements are chosen to be applied during “scarcity” hours when capacity is fully utilized. Assume that the system operator can implement a non-price rationing scheme (i.e. rolling-blackouts) when capacity constraints are reached in order to balance demand and sets a default price or price cap at $500/MWh under these conditions. Under these assumptions, an equilibrium in which generation suppliers can cover their total costs is characterized by less peaking capacity, less total capacity and nearly 200 hours of rolling blackouts each year, or 10 times more hours of rolling blackouts than in the efficient investment program for this example. The lower is the price cap the less investment will be forthcoming and the more hours of shortages requiring non-price rationing (rolling blackouts) will be necessary.

The absence of demand response from the system and/or constraints on the ability of the wholesale market to effectively integrate demand response into the formation of wholesale market prices creates the "missing money" problem and in the long run will lead to underinvestment in generating capacity and more hours of non-price rationing of demand.

4. THE MISSING MONEY PROBLEM IN PRACTICE

4.1 Causes of the Missing Money Problem

Joskow (2007) presents empirical evidence that the "missing money" problem is pervasive in organized wholesale markets in the U.S. and that it is a deterrent to generation investment consistent with the legally binding reliability standards that exist in the various regions of the U.S. What are the causes if this problem with the standard "energy only" market design? To answer this question, we will have to expand the simple models and examples that we have been using and incorporate considerations of the protocols that system operators utilize to meet engineering reliability standards when capacity is fully utilized or close to being fully utilized. These reliability standards and emergency operating protocols have been established historically by engineers, not economists, and have been carried over into liberalized markets from the old regime of regulated monopoly with little if any consideration given to how traditional network reliability criteria and behavioral protocols ---- rolling-blackouts, voltage reductions, network collapses, etc. --- might be integrated effectively into wholesale market mechanisms.

There are a number of wholesale market imperfections, regulatory constraints on prices, as well as the procedures system operators utilize to deal with operating reserve shortages that appear collectively to suppress spot market prices for energy and operating reserves below efficient prices during the small number of "scarcity" hours in a typical year when wholesale market prices should be very high.

a. Only a tiny fraction of electricity consumers and electricity demand during peak hours can see real time prices <u>and</u> can react quickly enough from the system operator's perspective to large sudden price spikes to keep supply and demand in balance consistent with operating reliability constraints. Neither the metering nor the control response equipment is in place except at a small number of locations.

b. In and of itself, the limited availability of real time meters and associated customer monitoring and response equipment is not a fatal problem, however. LSEs could enter into "priority rationing contracts" (Chao and Wilson (1987)) with retail consumers that would specify in advance the level of wholesale market prices at which customers would allow the system operator to implement demand curtailments. Retail customers entering into such contracts would receive a lower price per unit consumed on their standard meters (Joskow and Tirole 2006, 2007). They would not have to monitor real time prices themselves. This would be done (ultimately) by the system operator through a parallel contract with the retail consumer's LSE. However, priority rationing contracts require that the system operator can control the flows of power that go to individual customers and to have the capability to curtail individual customer demand on short notice. Except for the very largest customers, control over power flows does not go this far down into the distribution system and system operators can only curtail demand in relatively large "zones" composed of many customers (Joskow and Tirole 2006, 2007). That is, individual consumers cannot choose their individual preferred level of reliability when rolling blackouts are called by the system operator; their lights go off along with their neighbors' light.

c. System operators hold operating reserves for two reasons. One is to keep the probability of "controlled" non-price rationing of demand (rolling blackouts) low. The other is to keep the probability of an uncontrolled network collapse such as those that occurred in the Northeastern U.S. and in Italy in 2003 very low. Since the market also collapses in these situations, wholesale market prices are effectively zero and do not reflect consumer preferences to buy or generators' costs of supply. Individual consumers can do nothing to escape the consequences of a network collapse, aside from installing their own on-site generating facilities. Nor can individual generators profit from "scarcity" during a network collapse. As a result, there is no way for market mechanisms to fully capture the expected social costs of a network collapse. Joskow and Tirole (2007) argue that this gives operating reserves public good attributes. As a result, the efficient level of operating reserves will not be provided by market mechanisms but must be determined through some administrative process that reflects the probability and costs of a network collapse.

d. Rolling blackouts resulting from a shortage of generating capacity are extremely rare on electric power systems in developed countries.[10] Almost all of the "scarcity hours" are realized during operating reserve deficiency conditions when the system lies between the target level of operating reserves and the minimum level that triggers non-price rationing of demand. [11] Once price responsive demand has been exhausted, the price formation process during these conditions is extremely sensitive to small decisions made by the system operator and it is not evident that a simple market mechanism exists to produce the efficient price levels during these hours (Joskow and Tirole (2007)). Joskow (2007) gives two examples of actions taken by system operators during reserve deficiency conditions whose social costs are unlikely to be fully reflected in wholesale market prices.

1. The last thing that system operators typically do when there is an operating reserve deficiency prior to implementing rolling blackouts is to reduce system voltage by 5%. This reduces system demand and helps the system operator to keep operating reserves above the minimum level that would trigger rolling blackouts. However, reducing demand has the effect of reducing wholesale prices relative to their level at normal voltage and demand levels just as the system is approaching a non-price rationing state. Moreover, voltage reductions are not free. If they were free we could

[10] Almost all blackouts experienced by consumers result from equipment failures on the distribution network.

[11] The sequence of events and system operator behavior leading up to the rolling blackouts in Texas on April 17, 2006 provide an extremely informative insight into system operations during such scarcity conditions. Public Utility Commission of Texas (2006).

just operate the system at a lower voltage. Voltage reductions lead lights to dim, equipment to run less efficiently, on-site generators to turn themselves on, etc. These are costs that are widely dispersed among electricity consumers and are not reflected in market prices. Thus, the marginal social cost (in the aggregate) of voltage reductions is not reflected in market prices.

2. Markets for operating reserves typically define the relevant products (e.g. spinning reserves) fairly crudely. For example, spinning reserves may be defined as supplies from “idle” generating capacity that can be made available to the system operator within 10 minutes. The system operator may find it necessary to call on generating capacity that responds in, say, two minutes at particular locations on the network, to maintain the physical parameters of the network. The system operator typically has information about a more detailed set of generator characteristics than is reflected in product market definitions and can act upon this information when it thinks that it is necessary to do so to avoid rolling blackouts or a network collapse. When supplies from generators with more specific characteristics are needed by the system operator, it may rely on bilateral out-of-market (OOM) contracts to secure these supplies from specific generators and then dispatch the associated generating units as “must run” facilities at the bottom of the bid-stack. This behavior can inefficiently depress wholesale market prices received for energy and operating reserves by other suppliers in the market.

4.2 Market Power and Price Caps

The limited amount of real time demand response in the wholesale market leads to wholesale spot market demand that is extremely inelastic. Especially during high demand periods as capacity constraints are approached, this creates significant opportunities for suppliers to exercise unilateral market power. In the U.S., FERC has adopted a variety of general and locational price mitigation measures to respond to potential market power problems in spot markets for energy and operating reserves. These mitigation measures include general bid caps (e.g. $1000/MWh) applicable to all wholesale energy and operating reserve prices, location specific bid caps (e.g. marginal cost plus 10%), and other bid mitigation and supply obligation (e.g. must offer obligations) measures.

Unfortunately, the supply and demand conditions which should lead to high spot market prices in a well functioning *competitive* wholesale market (i.e. when there is true competitive "scarcity") are also the conditions when *market power* problems are likely to be most severe (as capacity constraints are approached in the presence of inelastic demand, suppliers' unilateral incentives and ability to increase prices above competitive levels, perhaps by creating contrived scarcity, increase). Accordingly, uniform price caps will almost inevitably "clip" some high prices that truly reflect competitive supply scarcity and consumer valuations for energy and reliability as they endeavor to constrain high prices that reflect market power. They may also fail to mitigate fully supra-competitive prices during other hours (Joskow and Tirole (2007)).

Many economists blame the missing money problem on these price caps alone. However, this argument is not consistent with the data on market prices in these wholesale markets. When one examines the full distribution of energy prices in these markets over a period of six years, it is evident that the price caps, despite being far below estimates of the VOLL, are rarely binding constraints (Joskow (2005), PJM (2006), New York ISO (2005), New England ISO (2005). Even during most "scarcity hours," market prices are below the price caps. Accordingly, it is unlikely that the price caps are the only source of the missing money problem. I believe that the effects (not the goal) of the other actions system operators utilize to maintain the operating reliability of the network play a much more important role in suppressing prices during scarcity conditions in the organized wholesale markets in the U.S. than do the price caps on energy and operating reserves.

5. IMPROVING "ENERGY ONLY" WHOLESALE MARKET PERFORMANCE

The fundamental source of the missing money problem is the failure of spot energy and operating reserve markets to perform in practice the way they are supposed to perform in theory. While I believe that the performance of spot wholesale energy markets can be improved, I do not believe that all of the problems, especially those associated with the implementation of engineering reliability rules and the associated behavior of system operators during scarcity conditions, can be fully resolved quickly if ever. Nevertheless, improving the behavior and performance of spot wholesale markets for energy and operating reserves can be a constructive component of a broader set of reforms. Joskow (2007) discusses a number of desirable wholesale market reforms that would contribute to reducing the magnitude of the missing money problem. They include:

1. Raising the price caps on energy and operating reserves during scarcity conditions: The $1000/MWh price cap in effect in most of the organized markets in the U.S. is a completely arbitrary number that is clearly below what the competitive market clearing price would be under most scarcity conditions. It would make sense to increase

the price caps to reflect reasonable values of VOLL if this action is combined with changes to the price formation process, more reliance on other approaches to mitigating market power, and continued reliance on market monitors as in all of the U.S. ISOs.

2. Require prices to rise automatically to the price cap when system operators take "out of market" (OOM) actions to deal with operating reserve deficiencies. Raising the price caps alone is not likely to help much if the price caps are not binding constraints. In order to make the higher price caps meaningful contributors to the missing money problem and to deal with the price formation problems that emerge when system operators implement reliability protocols when there are capacity constraints, it would make sense to adopt a rule that whenever a system operator issues a notice that operating reserve deficiency protocols will be implemented that the wholesale market prices for energy and operating reserves be moved immediately to the price cap. This is a rough and ready mechanism to get prices up closer to where they should be under scarcity conditions which responds to the challenges of implementing emergency response protocols, such as voltage reductions, while giving system operators the discretionary behavior that they may need to maintain network reliability and avoid network collapses.

3 Increase real time demand response resources: Increasing efforts to bring more demand response that meets the system operator's criteria for "counting on it" during scarcity conditions[12] can also help both to increase the efficiency with which capacity constraints are managed and improve the price formation process during scarcity

[12] This may require, for example, that demand respond to either price signals or requests for curtailment from the system operator within ten minutes or less. Demand response times has been identified as an issue in the investigation of the rolling blackouts in Texas (ERCOT) on April 17, 2006. See *Electric Transmission Week*, May 1, 2006, page 2, SNL Energy.

conditions. However, the ways in which demand response is brought into the system for these purposes is important. Demand response should be an active component of the price formation process and compete directly with resources on the supply side. The best way for this goal to be achieved is to structure demand response contracts as call contracts in which curtailments are contingent on wholesale prices rising to pre-specified levels.

4.. increase the number of operating reserve products sold in organized wholesale markets: Market performance would also be improved if market designs recognized that system operators need more refined "products" than are presently reflected in the ancillary service product definitions around which wholesale markets are now organized. For example, if the system operator needs "quick start" supply (or demand response) resources that can supply within five minutes rather than 10 minutes, it is better to define that as a separate product and to create a market for it that is fully integrated with related energy and ancillary service product markets rather than relying on out-of-market bilateral arrangements and "must run" scheduling at the bottom of the bid-based supply stack regardless of the marginal operating costs of these must-run units.

5. Review and adjust reliability rules and protocols*:* This leads to one final observation regarding the missing money problem that affects all proposed solutions to it. Many of the policy assessments of whether or not there is adequate investment in generating capacity turns on comparisons between market outcomes (investment in new and retirements of old generating capacity) and traditional engineering reliability criteria. These reliability criteria and associated operating protocols have been carried over from the old regime of regulated vertically integrated monopolies and may have reflected in

past efforts by vertically integrated utilities to justify excess generating capacity. It is not at all clear that even a perfectly functioning competitive wholesale market would yield levels of investment and reserve margins that are consistent with these reliability rules. Indeed, Cramton and Stoft's (2006, p. 33) observation that the capacity reserve margin criterion used in the Northeast reflects a VOLL of $267,000/MWh suggests that this reserve margin is much too high from the perspective of consumers' valuations for reliability. At the very least it would make sense to reevaluate these reliability criteria and to search for more market friendly mechanisms for achieving whatever reliability criteria are adopted.

6. USING FORWARD CAPACITY MARKET MECHANISMS TO CLOSE THE MISSING MONEY GAP

The reforms to wholesale energy markets discussed above should help to reduce the missing money problem associated with the operation of many "energy only" wholesale markets today. However, it is not at all obvious that the missing money problem will be completely solved with these reforms or that they can be implemented overnight. These reforms may also increase market power problems and further increase price volatility. I believe that reforms to spot markets need to be accompanied by a system of forward capacity obligations placed (ultimately) on Load Serving Entities (LSEs), the effective design of associated capacity markets and capacity payment mechanisms.[13] If properly designed, forward capacity markets can act as a safety valve to fill the net revenue gap that leads to the "missing money" problem. If these mechanisms are properly designed they can be consistent with the kinds of wholesale

[13] Crampton and Stoft 2006 and Joskow (2007) discuss why forward energy contracts alone will not solve the missing money problem.

market reforms discussed above and if these reforms are successful can be designed effectively to fade away over time.

6.1 A Two-State Model

Consider a simple case of the two-state example presented by Joskow and Tirole (2007). The low demand (off-peak) state has probability f_1 and the high demand (peak) state has probability f_2 . It is convenient to think about f_1 and f_2 as the number of hours of a typical year when the system experiences low demand and high demand respectively. There are two generating technologies; a base load technology with unit capital costs I_1 and marginal operating costs c_1 and a peaking technology with units capital costs $I_2 < I_1$ and marginal operating costs $c_2 > c_1$. Assuming that there is no non-price rationing, the efficient equilibrium prices that will support the optimal investment program will be:

Peak price: $p^*_2 = c_2 + I_2/f_2$

Off-peak price:[14] $p^*_1 = (I_1 – I_2)/f_1 + c_1 – f_2(c_2 – c_1)/f_1$

Clearly if the peaking capacity is paid p^*_2 for f_2 hours of the year, it will cover its capital and operating costs I_2 and f_2c_2. If the base load capacity is paid p^*_1 for f_1 hours of the year and p^*_2 for f_2 hours of the year, it will just cover its capital and operating costs I_1 and $(f_1+f_2) c_1$ as well. There is no missing money problem.

Joskow and Tirole (2007) produce the equivalent of the missing money problem by imposing a binding price cap on peak period prices $p_2^{max} < p_2$. In order to restore *investment* incentives a *capacity payment* $p_K = f_2(p^*_2 – p_2^{max})$ can be made to all capacity (peak and base load) that is utilized to meet demand during the peak period. In order to

[14] Because there are only two demand states and two generating technologies, this condition is more complicated than the simple pricing condition $p^*_1 = c_1$. In the numerical example above the points on the continuous load duration curve at which it becomes economical to switch from one technology to another have the property that $(I_1 – I_2)/f_1 = f_2(c_2 – c_1)/f_1$. In this case $p^*_1 = c_1$.

restore *consumption* or demand response incentives, the cost of making this capacity payment should be reflected in peak period prices so that peak period prices should now be $p_2 = p^{max}{}_2 + p_K/f_2$ per MWh (Joskow and Tirole (2007). Note that as the price cap rises toward the optimal peak period price $p^*{}_{2,}$ the capacity payment falls toward zero.

6.2 Capacity Payment Mechanism Design

The arithmetic of the appropriate capacity payment is fairly straightforward. However, designing the implementation mechanisms required to achieve the correct capacity payment is more complicated and involves harmonizing engineering reliability criteria with the developments of capacity markets to determine the appropriate capacity prices. The recently adopted capacity payment mechanisms in the U.S. typically start with the reliability criteria established by the responsible regional reliability organizations.[15] The primary generating capacity-related criterion is typically a generating capacity reserve margin measured by the difference between the expected system peak demand (D) before any curtailments and the peak generating capability (G) on the system assuming that all installed generating capacity is operating at the time of system peak. Qualifying demand response resources are in principle included in this generating capability number. The generating reserve margin criterion (R*) is then defined as R* = (G-D)/D and typically lies between 15% and 20% in the U.S. The target generating capability of the system is then G* = (1+R*)D.[16] Generating reserve criteria may be defined for the entire network controlled by the system operator and for individual sub-regions to reflect transmission constraints at the time of locational demand

[15] In the U.S. this organization would be the regional reliability council under which an SO operates and a national reliability organization provided for by the Energy Policy Act of 2005.

[16] In theory, the value for R* should reflect considerations of demand uncertainty, supply uncertainty, and the value of lost load from rolling blackouts and network collapses. In reality, the origins of these criteria are rather murky.

peaks (locational capacity prices). All retail load serving entities (LSEs) then have the obligation to pay for their proportionate share of this generating capacity/demand response obligation based on their own LSE load at the time of system peak.

LSEs can meet their capacity obligations either by contracting directly with generators for capacity to be available to supply energy at the time of system peak or by purchasing this capacity through an auction process conducted by the system operator. In the latter case, the system operator runs a series of auctions for qualifying generating capacity to meet the reliability criterion for installed generating capacity G.* The auction mechanism defines the price for generating capacity for one or more future periods. All LSEs are required to pay the market clearing price in the auction for their load-based share of the system generating capacity reserve obligation net of any generating capacity that they own or have contracted for separately outside of the auction ("self-supply"). [17] Generating units whose capacity clears in the capacity market and are counted by LSEs toward their capacity obligations have an obligation to offer energy to the wholesale spot market when requested to do so by the system operator or pay a significant performance penalty if they do not. Most of these proposals are structured as *forward* capacity obligations which require that capacity be auctioned three to five years into the future (Stoft and Cramton 2006). The multi-year forward capacity obligations are responses to concerns about the effects of price volatility on investment incentives and of market power on capacity prices.

[17] Self-supply can be easily accommodated by requiring generators with bilateral contracts to offer their capacity to the organized capacity market with a contract for differences with the LSEs with which they have pre-contracted and then including all LSE demand in the market as well. Effectively, the system operator buys capacity through the auction and bills LSEs for their share net of any self-supply by contract or ownership they have registered with the system operator prior to the auction.

6.3 Integrating Energy Markets with Capacity Payment Mechanisms

After the introduction of capacity obligations and associated capacity auction markets and capacity payments, the spot energy markets continue to operate as before, with whatever improvements discussed above may be introduced. Moreover, the market clearing price for capacity will reflect the attributes of the spot energy and ancillary services markets. Following the examples presented earlier, the equilibrium market clearing price (P_K) for generating capacity available during "scarcity" hours should equal the capital costs of a unit of peaking capacity (p_2) less the quasi-rents that a unit of peaking capacity is expect to earn (R_p) in the energy market during peak hours. The competitive capacity price $P_K = (P_2 - R_p)$ is then adjusted for expected forced outage rates and associated penalties (Joskow and Tirole (2007)).[18] Making capacity payments available in this way solves the missing money problem since the capacity price essentially acts as a safety valve to fill the gap between the capital costs of peaking capacity and the quasi-rents that an investor in peaking capacity must expect to earn in the energy and operating reserve markets to be willing to invest. Moreover, as the performance of the wholesale spot energy market improves during scarcity conditions, the expected quasi-rents produced for by the energy market for a unit of peaking capacity will rise toward $R_p = P_2$ and the capacity price P_K will fall toward zero. Thus, as the wholesale energy market's performance improves, capacity payments fade away.

6.4 Deficiencies of the Original Capacity Payment Mechanisms

Additional implementation details can be inferred from the performance problems associated with the first versions of capacity obligations, capacity markets and capacity

[18] Intermediate and base load capacity get the capacity price plus the quasi-rents they earn in the energy market consistent with the equilibrium conditions discussed above. In equilibrium all generating technologies that are included in the least cost portfolio cover their capital costs.

payments that were a feature of the organized wholesale electricity markets that began operating in the U.S. in the late 1990s[19]. The experience demonstrates that the implementation details are important because these early capacity payment mechanisms did not solved the missing money problem.

The original capacity payment mechanisms relied on cost-based calculations of deficiency payments that effectively placed a price cap on capacity prices. This cap kept realized capacity payments below the level necessary to make up for the net revenue gap realized on wholesale energy and operating reserve markets after liberalization. The new capacity payment mechanisms, on the other hand, retain a price cap to deal with potential market power problems, but the price cap is based on an analysis of the probability distributions of demand and supply so that on average the mechanism should yield a capacity price equal to P_2 before netting out any quasi-rents produced for a hypothetical peaker in the energy market. The proposed annual capacity price cap included in the forward capacity market adopted for the New England ISO is more than twice the old deficiency payment cap.

A second problem noted with the original capacity obligation/market systems is that they employed a hard value for the reserve margin and implied quantity of installed generating capacity (R^* and G^*) required to meet reliability criteria. This approach implied that the reliability value of generating capacity slightly above G^* was zero and that the value of any decrease in generation below G^* was effectively equal to the price cap. That is, the demand for capacity was equal to the price cap for $G < G^*$ and equal to zero for $G > G^*$. This led to very volatile capacity prices that jumped between close to

[19] These capacity obligations and deficiency payment rules were simply carried over from the tight power pools that existed in these areas prior to market liberalization. They were originally put in place to keep one regulated generating firm from free riding on the other members of the power pool.

zero and the price cap from year to year. The New York ISO has introduced a reserve demand curve that essentially smooths capacity prices around the target generating capacity reserve margin. The demand curve's structure is based on an assessment of the distribution of loss of load probabilities and the value of lost load. It is similar in concept to the capacity payment mechanism that was a component of the original wholesale market design in England and Wales prior to the introduction of NETA.

A third problem with the original capacity payment mechanisms was that the capacity market was effectively a short-term procurement market that did not give potential entrants an opportunity to participate in the auction, increasing the potential for incumbent generators to exercise market power in the capacity market as well as in the energy market. The new capacity obligations, capacity auction, and payment mechanisms in New England and PJM respond to this problem by turning the capacity auctions into forward markets for capacity that occur sufficiently far in advance of delivery that new entrants can participate in the auction. In New England, for example, the capacity auction will be for capacity that is to be available to the market over three years in the future.

A fourth problem with the original capacity market arrangements was that they provided investors considering entering the market with no way of locking in capacity prices for any time period in advance of completion. Whether this concern reflects uncertainty per se or potential regulatory hold-up problems is unclear. However, the New England forward capacity market and payment mechanism allows new entrants to lock in capacity prices determined in the auction for a period of up to five years after the forward capacity delivery date at their choice.

A fifth problem with the original capacity obligation/market arrangements was that generators had poor incentives to be available during hours when capacity is constrained because capacity payments were not tied to actual performance during these hours but rather to historical availability experience. This problem is exacerbated by the failure of energy prices to rise to high enough levels during these critical periods. The new mechanisms include penalties for generators who are not available to perform when they are most needed.

A sixth problem with the original arrangements (and the primary initial motivation for the reforms in New England and PJM) was that capacity obligations were applied for the system operator’s entire network and did not reflect transmission congestion and local reliability and associated installed capacity criteria . At first blush, this problem may seem a little surprising since the Eastern and Midwestern markets in the U.S. rely on locational marginal price (LMP) mechanisms for energy that yield prices that are supposed to reflect congestion (Joskow 2006). However, the same market and institutional failures that suppress energy prices generally, also affect prices in constrained areas. To respond to this problem, the new capacity market mechanisms allow for capacity obligations and capacity prices to be determined for sub-regions where there are congestion problems (e.g. Southwestern Connecticut, New York City, Northern New Jersey.)

6.5 <u>Market Power Mitigation Considerations</u>

A final criticism of the original capacity market arrangements is that they failed to do anything about market power in the energy market or to stimulate more hedging of energy price volatility for retail customers (‘hedging load”). The forward capacity market mechanism approved for New England contains an interesting component that responds to these concerns. Each year the system operator will calculate the quasi-rents earned by a hypothetical peaking unit for sales of energy and operating reserves in the spot market ("Peak Energy Rents" or "PER") and deduct these rents from the capacity price determined in the auction. The PER is calculated based on a strike price for the hypothetical peaking unit that is equal to its marginal generation cost.

This provision has several effects. First, it hedges load against peak period energy price spikes since as peak period prices increase in the energy market the net price they end up paying for capacity per se decreases. Many consumers appear to value this type of hedge. Second, it provides a full net revenue hedge to peaking capacity that performs as expected and a partial hedge to base load and intermediate capacity. This responds to the argument that more price certainty is necessary to attract investment with lower rate of return expectations. Third, it reduces incentives to exercise market power in the energy market since higher spot market prices do not benefit generators that are fully hedged in this way. Finally, it provides good performance incentives. A generator that does not meet the performance targets and parameters used to calculate PER for a hypothetical peaker will lose money on the PER adjustment (as well as from other performance incentives). A peaker that can realize better performance keeps the additional net revenues.

6.6 Demand-Side Considerations

Most of the discussion of capacity payment mechanisms has focused on the supply side. I also want to emphasize a point that I made earlier. To fully restore appropriate incentives to market participants, the demand side of the market should be treated symmetrically with the supply side. Demand response resources that are compatible with the system operator's reliability criteria should be compensated at levels equivalent to what is paid to generators to make capacity available during capacity constrained periods. Moreover, the price paid for capacity should ideally be reflected in prices paid by wholesale markets and retail consumers during these same critical periods. This should be a goal of further refinements in the forward capacity market framework.

7. CONCLUSION

Policymakers in many countries are concerned that competitive wholesale markets for electricity do not provide adequate incentives for investment in sufficient quantities of generating capacity or an efficient mix of generating capacity consistent with acceptable reliability criteria. These concerns are creating barriers to full implementation of efficient electricity sector liberalization. There is now extensive empirical evidence that these concerns are valid, at least in some wholesale markets, so they cannot be easily dismissed. One important source of the problem is the failure of wholesale spot markets for energy and operating reserves to produce prices for energy during periods when capacity is fully utilized to meet the demand for energy and operating reserves that are high enough to support investment in an efficient (least cost) portfolio of generating capacity. This is the so-called "missing money" problem. There are a number of reforms that can be made to "energy only" wholesale markets to reduce the magnitude of the missing money problem. However, these reforms will take time to implement fully and it is far from obvious that market mechanisms can be designed to incorporate the social costs of all reliability actions taken by system operators into market prices. Electricity sector liberalization may not survive a period of underinvestment, increased hours of rolling blackouts, and higher probabilities of network collapses.

A set of forward capacity obligation, capacity market, and capacity payment mechanisms can be implemented, at least as transitional mechanisms, to mitigate the missing money problem. These mechanisms can be designed to be compatible with improvements in the efficiency of spot wholesale markets, the continued evolution of competitive retail markets, as well as to restore incentives for efficient investment in generating capacity and demand response consistent with operating reliability criteria applied by system operators. Capacity obligation and payment mechanisms can also be designed to respond to investment disincentives that have been associated with volatility in wholesale energy prices by hedging energy prices during peak periods as well as responding to concerns about regulatory opportunism by establishing forward prices for capacity for a period of up to five years. These hedging arrangements also reduce the incentives of suppliers to exercise market power.

TABLE 1

HYPOTHETICAL ELECTRIC GENERATION SYSTEM WITH DEMAND RESPONSE “TECHNOLOGY”

Generation Technology	Annualized Capital Costs	Operating Costs
	$/MW/Year	$/MWH
Base load	$240,000	$20
Intermediate	$160,000	$35
Peaking	$ 80,000	$80
Demand response (VOLL)	-0-	$4000

Load Duration Curve (See Figure 1)

$$D = 22{,}000 - 1.37H \quad [0 < H < 8760]$$

D = System load

H = Number of hours system load reaches a level D

Source: Joskow (2007)

TABLE 2

LEAST COST MIX OF GENERATING TECHNOLOGIES AND RUNNING TIMES FOR HYPOTHETICAL SYSTEM WITH DEMAND RESPONSE

Generating Technology	Capacity (MW)	Running hours	Total Cost ($billions)
Base load	14,694	5333 – 8760	$5.940
Intermediate	4,871	1778 – 5333	$1.385
Peaking	2,407	20.4 – 1778	$0.3657
Demand Response	28	0 – 20.4	$0.0011
TOTAL	22,000		$7.692

Source: Joskow (2007)

TABLE 3

SHORT-RUN MARGINAL COST + SCARCITY PRICING
PRICE DURATION SCHEDULE

Marginal Technology	Short-run Marginal Cost $/MWh	Duration hours
Base load	$20	3427
Intermediate	$35	3556
Peaking	$80	1757
"Scarcity" (Demand Response)	$4000	20

Source: Joskow (2007)

TABLE 4

PROFITABILITY OF SHORT-RUN MARGINAL COST + "SCARCITY" PRICING OF ENERGY PRODUCTION FOR LEAST COST SYSTEM

Generating	Revenues	Total Cost	Shortfall	
Technology	($billions)	($billons)	$(billions)	$/MW/Year
Base load	$5.940	$5.940	-0-	-0-
Intermediate	$1.385	$1.385	-0-	-0-
Peaking	$0.366	$0.366	-0-	-0-
Demand Response	$0.0114	$0.0114	-0-	-0-

Source: Joskow (2007)

TABLE 5

QUASI-RENT DISTRIBUTION WITH MARGINAL COST + "SCARCITY" PRICING FOR HYPOTHETICAL LEAST COST SYSTEM

	Net Revenues Earned	
Technology	Marginal Cost Pricing Hours	Scarcity Pricing Hours
Base load	67%	33%
Intermediate	50%	50%
Peaking	0%	100%

Source: Joskow (2007)

TABLE 6

PROFITABILITY OF THE LEAST COST SYSTEM WITH SHORT-RUN MARGINAL COST PRICING OF ENERGY PRODUCTION

Generating	Revenues	Total Cost	Net Revenue Shortfall	
Technology	($billions)	($billons)	$(billions)	$/MW/Year
Base load	$4.765	$5.940	($1.176)	$80,000
Intermediate	$0.996	$1.385	($0.390)	$80,000
Peaking	$0.173	$0.368	($0.195)	$80,000
Demand Response	--- Non-price rationing -----			
	$5.934	$7.694	($1.760)	

Source: Joskow (2007)

REFERENCES

Boiteux, M. (1960), “Peak Load Pricing,” *Journal of Business*, 33:157-79 [translated from the original in French published in 1951.]

Boiteux, M. (1964), “The Choice of Plant and Equipment for the Production of Electric Energy,” in James Nelson, ed. *Marginal Cost Pricing in Practice*, Englewood Cliffs, N.J., Prentice-Hall.

Chao, H.P. and R. Wilson (1987), “Priority Service: Pricing, Investment and Market Organization,” *American Economic Review*, 77: 89-116.

Cramton, Peter and Steve Stoft (2006), “The Convergence of Market Designs for Adequate Generating Capacity,” manuscript, April, 25, 2006.

Crew, Michael A. and Paul R. Kleinforder (1976), “Peak Load Pricing with Diverse Technology,” *Bell Journal of Economics*, 7(1): 207-231.

Gilbert, R., K. Neuhoff, and D. Newbery, 2004. “Allocating Transmission to Mitigate Market Power in Electricity Networks,” *Rand Journal of Economics*, 35(4), 691-709.

Green, R (2004). “Electricity Transmission Pricing: How Much Does it Cost to Get it Wrong?” MIT CEEPR Working Paper 04-020WP, September. http://web.mit.edu/ceepr/www/2004-020.pdf

ISO New England (2006), FERC Filing on Proposed LICAP Settlement, http://www.pjm.com/markets/market-monitor/som.html

Joskow, P.L. (1976), “Contributions to the Theory of Marginal Cost Pricing,” *Bell Journal of Economics*, 7(1), 197-206.

Joskow, P.L. (2005). “The Difficult Transition to Competitive Electricity Markets in the United States.” *Electricity Deregulation: Where To From Here?*. (J. Griffin and S. Puller, eds.) , Chicago, University of Chicago Press.

Joskow, P.L. (2006), “Markets for Power in the U.S.: An Interim Assessment,” *The Energy Journal*, 27(1): 1-36.

Joskow, P.L. (2007), “Competitive Electricity Markets and Investment in New Generating Capacity,” in *The New Energy Paradigm*, Dieter Helm, ed., Oxford University Press, 2007.

Joskow P.L. and J. Tirole (2000), “Transmission Rights and Market Power on Electric Power Networks,” *Rand Journal of Economics*, 31(3), 450-487.

Joskow, P.L. and J. Tirole (2006), “Retail Electricity Competition,” *Rand Journal of Economics*, 37(4), 799-815.

Joskow, P.L. and J. Tirole (2007), “Reliability and Competitive Electricity Markets,” *Rand Journal of Economics*,38(1), 60-84.

New York ISO (2005), “2004 State of the Markets Report,” prepared by David Patton. July, http://www.nyiso.com/public/webdocs/documents/market_advisor_reports/2004_patton_final_report.pdf

PJM Interconnection (2006), *2005 State of the Market Report*, http://www.pjm.com/markets/market-monitor/som.html

Public Utility Commission of Texas (2006), *Investigation into the April 17, 2006 Rolling Blackouts in the Electric Reliability Council of Texas Region: Preliminary Report*, April 17, 2006.

Sioshansi, F.P. and W. Pfaffenberger (2006), *Electricity Market Reform: An International Perspective*, Elsevier

Stoft, Steven (2002), *Power System Economics*, IEEE Press.

Turvey, R. (1968), *Optimal Pricing and Investment in Electricity Supply: An Essay in Applied Welfare Economics*, Cambridge, MA: MIT Press.