COMPETITIVE ELECTRICITY MARKETS AND INVESTMENT IN NEW GENERATING CAPACITY

Paul L. Joskow[1]
MIT
June 12, 2006

INTRODUCTION

Policymakers in many countries are expressing concerns that competitive wholesale electricity markets are not providing appropriate incentives to stimulate "adequate" investment in new generating capacity at the right time, in the right places, and using the right technologies. These concerns are often expressed in the context of concerns about "supply security," "reliability," "resource adequacy," or "supply diversity." In most cases the concerns have been raised as policymakers observe growing electricity demand, shrinking reserve margins and rising prices but little evidence of investment in new generating capacity responding to balance supply and demand consistent with traditional metrics for generation resource "adequacy." Many economists and market enthusiasts dismiss these concerns as reflecting the misguided conclusions of nervous politicians and system engineers who do not understand how markets work and who have not made the intellectual transition to a world of liberalized electricity markets. Nevertheless, there are a growing number of recent situations in which state-owned entities have stepped in to contract for additional generating capacity or where policymakers have required incumbent distribution companies to contract for new

[1] This paper builds on research discussed in Joskow (2005), Joskow (2006) and Joskow and Tirole (2005a). I have benefited enormously from conversations with Jean Tirole and Steve Stoft about these issues and, in particular, the recent paper by Cramton and Stoft (2006) provides an excellent technical discussion of several of the issues discussed here. Financial support from the MIT Center for Energy and Environmental Research is gratefully acknowledged. The views expressed here are my own and do not necessarily reflect the views of those with whom I have worked on these issues or of MIT.

supplies to mitigate resource adequacy concerns (e.g. Chile, Brazil, New Zealand, Ontario, California).[2]

In this paper I will argue that, at least based on U.S. experience with organized competitive wholesale power markets for electric energy and operating reserves, there are a number of market imperfections and institutional constraints that have the effect of keeping wholesale prices for energy and operating reserves below their efficient levels during hours when prices should be very high and provide insufficient net revenues to support the capital costs of an efficient portfolio of generating facilities. If this situation is allowed to persist it will in turn lead to underinvestment in generating capacity and to higher rates of power supply emergencies and involuntary rationing (blackouts). These problems have been exacerbated in the U.S. by instability in the wholesale market designs and market rules that characterize these wholesale markets (continuing reforms of the reforms), uncertain commitments by government policymakers to liberalization (calls for re-regulation), and an incomplete transition to a stable retail competition framework. At least some of these problems are likely to characterize competitive electricity markets in some other countries. That's the bad news. The good news is that these problems can be fixed with appropriate reforms to wholesale and retail market designs and credible government commitments to market liberalization.

The concerns about investment in new generating capacity reflect one or more of several interrelated groups of real or imagined problems with competitive wholesale electricity markets. First, it has been argued that competitive wholesale electricity markets for energy and operating reserves do not and perhaps cannot credibly provide sufficient net revenues to attract adequate investment in generation to meet conventional

[2] For example see the chapters in Sioshansi and Pfaffenberger (2006).

operating and investment economic efficiency and reliability criteria. According to this view, spot wholesale electricity market prices for energy and operating reserves will simply not be high enough to cover both the operating costs and the capital investment costs (including an appropriate risk adjusted cost of capital) required to attract new investment in long-lived generating capacity to support a least cost generation supply portfolio consistent with mandatory reliability criteria. Wholesale spot market prices in turn are reflected in forward prices for power that are too low as well through the normal operation of inter-temporal arbitrage behavior. I will follow Cramton and Stoft (2006) and refer to this as the "missing money" problem.

Second, 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 on 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.

Finally, it is 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. As I will discuss in more detail below, 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.

In this paper I will focus on the first set of concerns --- what Cramton and Stoft (2006) call the missing money problem ---, discuss empirical evidence indicating that it is a real problem in the organized wholesale power markets in the U.S., and identify its causes. I do not think too much of the argument that price uncertainty per se deters investment, though I will discuss how restrictions on the natural evolution of retail market institutions can contribute to a failure of normal market-based risk allocation mechanism to operate properly. The issues related to investment disincentives caused by opportunism or hold-up concerns are real and require more attention. While mandating that retail suppliers enter into long term contracts may be a solution to this problem from the perspective of investors (Joskow (1987)), it is a solution that is not compatible with

the effective diffusion of retail electricity competition and may deter further improvements in wholesale market institutions. Finally, I discuss a series of reforms built around (a) improvements in spot wholesale energy markets and (b) the introduction of forward capacity markets with particular attributes that can resolve most of the problems that have been identified and are compatible with the continued evolution of a healthy retail competition framework.

BACKGROUND

Questions have been raised about whether competitive wholesale and retail markets for power would produce adequate generating capacity investment incentives to balance supply and demand efficiently since the transition to competitive electricity markets began. Until 2001, the wholesale market system in England and Wales provided for additional capacity payments to be made to all generators scheduled to supply during hours when supply was unusually tight (i.e. when the loss-of-load probability was relatively high).[3] The wholesale markets created and managed by the Eastern Independent System Operators (ISOs) in the U.S. during the late 1990s have continued their traditional policies of requiring distribution companies (or more generally "load serving entities" or "LSEs" to encompass competitive retail electricity suppliers) to enter into contracts for capacity to meet their projected peak demand plus an administratively determined reserve margin. Argentina's competitive electricity market system also included capacity payments to stimulate investment in reserve generating capacity. In Chile, distribution utilities are required to enter into forward contracts to meet forecast demand plus a reserve margin. The system in Columbia also imposes capacity

[3] This payment mechanism was dropped when the New Electricity Trading Arrangements (NETA) system was introduced in 2001.

obligations. California's wholesale electricity market design did not impose capacity, reserve or forward contract obligations and the California electricity crisis of 2000-2001 is sometimes (erroneously) blamed on underinvestment in generating capacity. Capacity obligations are now being introduced in California in the form of generating reserve margin criteria and forward contracting obligations. On the other hand, the wholesale market in England and Wales abandoned capacity payments when the New Electricity Trading Arrangements (NETA) where introduced and Texas (ERCOT) has never had capacity payments or capacity obligations. There appears to be no interest in introducing them in either market.

Questions about whether wholesale markets will bring forth adequate investments in generating capacity arises naturally from the unusual characteristics of electricity supply and demand: (a) large variations in demand over the course of a year; (b) non-storability; (c) the need to physically balance supply and demand at every point on the network continuously to meet physical constraints on voltage, frequency, and stability; (d) the inability to control power flows to most individual consumers; (e) limited use of real time pricing by retail consumers, and (f) that even under the best of circumstances (i.e. with effective real time pricing of energy and operating reserves) non-price mechanisms (blackouts) will have to be relied upon from time to time to ration imbalances between supply and demand to meet physical operating reliability criteria because markets cannot clear fast enough to do so.[4]

[4] In response to questions about why demand response was not relied upon to respond to the sudden loss of 1,100 Mw of generating capacity that led to rolling blackouts in Texas on April 17, 2006, a representative of the ISO is reported to have said: "In this case, when four generators tripped, it was just bang-bang-bang-bang." *Electric Transmission Week*, April 24, 2006, pages 1 and 12, SNL Financial LC.

These attributes have a number of implications. First, a large amount of generating capacity that is available to meet peak demand plus the associated operating reserve requirements supplies relatively small amounts of energy during the year. For example, in New England in 2001, 93% of the energy was supplied by 55% of the installed generating capacity while the remaining 45% of the capacity supplied only about 7% of the energy.[5] Potential investors in new generating capacity must expect to cover their variable operating costs, their fixed operating and maintenance costs, and their capital costs from sales of energy and operating reserves over the life of generating capacity under consideration. The return of and on the associated capital investment in new generating capacity is the difference between the prices they receive for generation services (including capacity payments, if any) and their operating (primarily fuel) costs – what I refer to here as "net revenues." The profitability of generating units that are likely to operate only for a relatively small number of hours in each year ("peaking capacity") is especially sensitive to the level of prices that are realized during the small number of high demand hours in which they provide energy or operating reserves.

Second, the generating capacity available to supply energy at any point in time must always be greater than the demand for energy at that point in time as a result of the physical need to carry "inventory" in the form of generators providing frequency regulation and operating reserve services. That is, generating capacity (or in principle demand response) must be available that is either "spinning" or available to start up quickly to provide energy to balance supply and demand at each location on the network in response to real time variations in demand and unplanned equipment outages. When

[5] Sithe Energy presentation, IAEE, Boston Chapter, February 19, 2003.

these operating reserves fall below a certain level because available generating capacity and demand response resources are fully utilized (e.g. 7% of demand), system operators begin to take actions to reduce demand administratively according to a pre-specified hierarchy of “operating reserve shortage” actions. The final actions in this hierarchy are voltage reductions and non-price rationing of demand (rolling blackouts). I will discuss system operator behavior during such "scarcity" or "operating reserve shortage" conditions in more detail below as they play a central role in explaining the missing money problem in the U.S.

Finally, limited reliance on real time pricing, the inability to control real time power flows to all but the largest retail consumers, the potential for and economic attributes of a network collapse, the attributes of system operating protocols such as voltage reductions, undermine the ability of market mechanisms alone to choose the efficient level of system reliability. Reliability has public goods attributes and we cannot expect the market to provide the efficient level of reliability. Whether the market can choose the efficient level of reliability or not, a variety of administrative reliability rules and operating protocols have been carried over from the old regime of vertically integrated monopoly to the world of liberalized electricity markets. These reliability rules have important implications for market behavior and performance and about assessments of the “adequacy” of investment in generating capacity and the associated probability of rolling blackouts and network collapses.

WHOLESALE ELECTRICITY MARKET BEHAVIOR AND PERFORMANCE IN THEORY

To oversimplify for expositional purposes, a well functioning perfectly competitive wholesale electricity market will operate in one of two states of nature. Under typical operating conditions (State 1), market clearing prices for energy and operating reserves should equal the marginal (opportunity) cost of the last increment of generating capacity that just clears supply and demand at each point in time. In the case of wholesale electric energy supply, this price is the marginal cost of producing a little more or a little less energy from the generating unit on the margin in the bid-based merit order. Figure 1 depicts the spot market demand for electricity and the competitive supply curve for electricity under typical operating conditions (State 1). Inframarginal generating units earn net revenues or quasi-rents that contribute to the recovery of their fixed operating and capital costs whenever the market clearing price exceeds their own marginal generation costs. In the case of operating reserves, the efficient price is (roughly) equal to the difference between the price of energy and the marginal cost of the next increment of generation that could supply energy profitably if the price of energy were slightly higher plus any direct costs incurred to provide operating reserves (e.g. costs associated with spinning). This price for operating reserves is equal to the marginal opportunity cost incurred by generators standing in reserve rather than supplying energy. Under typical operating conditions (State 1) the price of operating reserves will be very small --- close to zero, and far below the price of energy.

The second wholesale market state (State 2) is associated with a relatively small number of hours each year when there would be excess demand at a wholesale price that is equal to the marginal production cost of the last increment of generating capacity that

can physically be made available on the network to supply energy plus operating reserves. In this case, the market must be cleared "on the demand side." That is, consumers bidding to obtain energy would bid prices up to a (much) higher level reflecting the value that consumers place on consuming less electricity as demand is reduced to match the limited supplies available to the market (or value of lost energy or load -- VOLL). This second state is depicted in Figure 2. In Figure 2, the area labeled R_{mc} represents the quasi-rents that would be earned by infra-marginal generators if the wholesale price is equal to the marginal generating cost of the least efficient generator on the system required to clear the market. The area labeled R_s reflects the additional scarcity rents from allowing prices to rise high enough to ration scarce capacity on the demand side to balance supply and demand. In what follows, I will refer to the conditions depicted in Figure 2 as competitive "scarcity" or "shortage" conditions.[6]

Under competitive scarcity conditions the competitive market clearing price of energy will now generally be much higher than the marginal production cost of supplying the last available increment of energy from generating capacity available to the network, reflecting the high opportunity cost (value of lost energy or lost load – VOLL in what follows) that consumers place on reducing consumption by a significant amount on very short notice. Furthermore, while the price of operating reserves will continue to be equal to the marginal opportunity cost incurred by generators standing in reserve rather than supplying energy, the opportunity cost of standing in reserve rather than supplying energy will rise significantly as well in response to the higher "scarcity value" of energy. All generating units actually supplying energy and operating reserves in the spot market

[6] To distinguish it from contrived scarcity resulting from suppliers withholding supplies from the market to drive up prices.

during scarcity conditions would earn substantial "scarcity rents." These scarcity rents in turn help to cover the fixed capital and operating costs of all generating facilities.

In a hypothetical well functioning competitive electricity market, and in particular ignoring the market imperfections associated with the market-provision of reliability, price signals for energy bought and sold in the market not only induce the right amount of generating capacity (and associated levels of reliability), but also the right mix of generating technologies. Because electricity is non-storable and demand varies widely over the course of a year, the most economical portfolio of generating plants will include technologies with a variety of capital cost/operating cost ratios. Base load generating facilities (typically nuclear or coal) have relatively high capital costs and low operating costs. These facilities are economical to build if it is efficient to operate them for a large fraction of the hours of each year. Intermediate load facilities (typically gas or oil fueled) have lower capital costs and higher operating costs than base load facilities. These facilities typically operate for 20% to 50% of the hours during the year. Finally, peaking facilities have the lowest capital costs and the highest operating costs per unit of capacity. These facilities are expected to be economical to operate from a few hours per year up to (say) 20% of the hours during the year.[7]

For base load and cycling units, the net revenues they earn during scarcity conditions may account for a significant fraction of the total net revenues they earn throughout the year. For peaking capacity that supplies energy or operating reserves primarily during such scarcity conditions, the net revenues they earn during these periods

[7] There does not exist a distribution of generating technologies that reflect a continuum of capital/operating cost ratios. However, the more options there are along such a distribution the better can be the match between generating technologies and the number of hours they will operate each hear to meet demand.

will account for substantially all of the net revenues available to cover their fixed costs (capital, maintenance and operating.). The number of hours in which "scarcity" conditions emerge depends upon the amount of generating capacity that has been installed and is physically available to operate relative to the tail of the distribution of aggregate demand realizations during the year. The quantity and type of generating capacity that is physically available to the network in a market context will then depend on investors in generating capacity balancing the costs of additional investments against the net revenues they expect to receive, including the "scarcity" rents produced under State 2 conditions, from spot market sales and through sales pursuant to forward contracts if suppliers choose to hedge market prices risks. The prices for such forward contracts are necessarily linked directly to expected wholesale spot market prices for energy through intertemporal arbitrage and consumer and supplier preferences for market price risk.

This simple theoretical analysis of a well-performing wholesale market has so far largely ignored uncertainty. Uncertainty enters short run operating (dispatch) behavior and long run investment behavior in a number of ways. Electricity demand is uncertain in both the long run and the short run. From a long run investment perspective, electricity demand depends on the average level of future electricity prices, the prices of substitute fuels, the replacement rates of appliances and equipment and both the level and composition of aggregate economic activity. In the short run, given the stock of appliances and equipment, electricity demand is particularly sensitive to weather conditions since weather variations lead to large variations in heating and cooling demand. Short run price and income elasticities are very low. On the supply side, from

an investment perspective, there is uncertainty about future electricity prices, fuel prices and the rates of entry new and exit of old generating capacity. In the short run, there is uncertainty about unplanned outages of generating facilities and spot prices, reflecting the interactions of uncertain demand and uncertain supply. Uncertainty on the supply and demand side introduces volatility into spot prices over and above the natural variability in prices associated with variable demand and differences in the short run marginal costs of operating diverse generating technologies. It will also lead to a least cost investment portfolio that will have more nominal generating capacity (measured before taking account of forced outage rates) than the expected (mean) level of peak demand. The difference between the nominal generating capacity on the system and the expected peak demand is the system's expected "reserve margin."

Historically, when the electricity sector was composed of vertically integrated regulated monopolies, these aspects of uncertainty affected investment and operating decisions in important ways. From an investment perspective, long-term planning protocols reflected longer term uncertainty on the supply and demand sides by establishing target "reserve margins" over and above the expected level of peak electricity demand. These reserve margins were based on forecast levels of peak demand and forecasts levels of capacity, assuming that all of the capacity would be available at the time of system peak. So, for example, in the U.S., systems were planned to yield an average "planned" reserve margin of 15% to 20%. The reserve margin typically could include contracted demand response that the system operator could control but did not assume that demand would otherwise respond to rapid changes in real-time prices.

From a short run operating perspective, the quantity of generating capacity scheduled to be available to supply electrical energy includes capacity used for frequency regulation, operating reserves and replacement reserves. In a typical system these "operating reserves" account for an additional 10% to 12% of generating capacity above the actual demand for energy at any particular time. Generating capacity is scheduled in this way as a result of the perceived need to have "quick response" generation resources available to respond to short-term fluctuations in demand and unplanned outages of generating and transmission capacity, in order to keep the probability of non-price rationing (rolling blackouts) and cascading network outages (network collapse) very low. These operating and investment criteria are typically enshrined in various engineering reliability rules that have been carried over without much if any changes into the world of liberalized electricity markets.

The role of operating reserves in real electricity systems changes the static notion of a capacity constraint in real time operations as typically reflected in simple market models (e.g. as in Figure 2). Capacity constraints are now "soft" constraints that exceed the actual demand for energy on the system at any particular time. In normal operations, the generating capacity scheduled by the system operator to supply energy quickly through the wholesale energy and operating reserve markets will include about 10% operating reserves of one type or another over and above the demand for energy. When this target level of operating reserves cannot be maintained because there is no additional generating capacity or demand response available for the system operator to call upon, an "operating reserve emergency" or "operating reserve shortage" will be declared. That is, the capacity constraint is effectively reached when the generating capacity available to

the system operator falls below (say) 110% of current demand for energy (or forecast demand for the next few hours). Accordingly, a more realistic characterization of capacity constraints (State 2 conditions) depicted in Figure 2 should include operating reserves in total capacity required to meet any given level of demand. Moreover, as I will discuss in more detail presently, the "soft" capacity constraint created by the operating reserve targets and system operator reliability protocols in the face of operating reserve constraints significantly complicates the price formation process during scarcity conditions.

NUMERICAL EXAMPLES[8]

The simple economics of the efficient utilization, investment and pricing for an electric generating system is usefully clarified with a couple of simple numerical examples that, for simplicity, ignore uncertainty and public goods aspects of reliability. Table 1 displays the parameters of three hypothetical electric generating technologies with different capital cost/operating cost ratios and a hypothetical load duration curve representing the number of hours during the year the aggregate system demand or "load" reaches any particular demand level.[9] The capital costs of a generating facility are fixed costs once the investment to build it has been made. The operating costs vary directly

[8] 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.

[9] The arithmetic associated with the example is in continuous time though for simplicity I will refer to "hours" of load duration and generator utilization in "hours" in the discussion.

with the production of electrical energy from the generating facility.[10] 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. Finally, there is a "peaking" technology with still lower capital costs and higher operating costs. In the example, demand is equal to 10,000Mw 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. For now, we will assume that the annual hourly system demand profile summarized in the load duration curve is not sensitive to prices. This assumption will be relaxed presently. We also ignore uncertainty on the demand side and the supply side for now.

Total costs (capital plus operating) per unit of generating capacity vary 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 upon how many hours each year it is anticipated that each will be economical to "dispatch" to supply electricity. If a generating unit is expected to operate economically 8760 hours per year, the base load technology is the lowest cost choice. If a unit of generating capacity is expected to be economical to run, for example, only 4,000 hours per year, then intermediate load technology is the lowest cost choice. If the capacity is expected to be economical to run, for example, 200 hours per year, then peaking technology is the least cost option. These relationships for this numerical example are depicted in the top panel of Figure 3. The top panel yields the duration of demand at which each technology is economical from a total cost (capital plus operating)

[10] 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.

perspective. The lowest cost mix of investments in generating technology can then be determined by "fitting" the total cost of building and operating each generating technology at alternative utilization rates to the load duration curve for the system (since electricity cannot be stored). This can be accomplished graphically by including the load duration curve in the bottom panel in Figure 3 and matching the technology in the top panel to the load duration in the bottom panel at which it is the least cost technology. The quantity of capacity of each technology that makes up the least cost generating investment portfolio can then be read off of the vertical axis in the bottom panel of Figure 3 at the load duration cutoff points for each technology.

For this example, Table 2 displays the least costs mix of generating capacity, 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 parameters in Table 1 and the graphical representation in Figure 3. In this example, the least cost mix includes a lot of base load capacity, a much smaller amount of intermediate capacity and an even smaller amount of peaking capacity.

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

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 market will operate in a regime where there is "excess capacity" as in Figure 1 (State 1) above. In a perfectly competitive market, prices will reflect short run marginal operating costs under these "State 1" conditions. 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, peaking capacity is marginal and the perfectly competitive market price will be $80/Mwh up to the point where capacity is fully utilized. Table 3 displays the number of hours that each technology is the marginal supplier. Let me defer for now a discussion of what the price would be when demand and capacity are both exactly 22,000 Mw in this example.

Table 4 displays the revenues, total costs and difference between revenues and total costs (shortfall or net revenue gap) for each technology and in the aggregate under the short run marginal cost pricing scenario just discussed. 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 <u>at the efficient investment levels.</u> The shortfall turns out to be $80,000/Mw of installed capacity for all technologies. 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 revenue of 30%). Note for future reference that the required $80,000/Mw of generating capacity is also exactly equal to the annualized capital charges for a Mw of peaking capacity. Clearly either some type of "capacity" charge equal to the capital cost of a peaking turbine must for paid for each unit of capacity used at the time of system peak when capacity is fully utilized (effectively $80,000 per peak Mwh consumed when demand is 22,000 Mw) or some alternative market mechanism must emerge to increase energy prices significantly during some hours of the year.

To capture how (simplified) well functioning competitive wholesale energy markets are supposed to function we must introduce some demand elasticity into the example. It is convenient for the exposition here, and to capture the way sytem operators think about demand response, to conceptualize "demand response" as a technology option through which demand is paid to reduce consumption. The payments reflect the marginal value consumers place on consuming less energy in the very short run --- what is generally referred to as the "value of lost load" or VOLL (See Stoft (2002), Chapter 2-5). Accordingly, I expand the numerical example to include an additional demand response technology which reflects a VOLL of $4000/Mwh. Table 5 expands the example reflected in Table 1 to include this fourth "demand response" technology with a VOLL of $4000/Mwh. As I will discuss presently, this value for VOLL is well within the range of available estimates used in practical applications (e.g. in the old E&W pool and in Australia).

We can now derive the least cost mix of the four "generating technologies," including demand response. The result is displayed in Table 6 which should be compared to Table 2. With the demand response option available, 28 Mw of demand response is

substituted for peaking capacity and demand with durations of between one second to 20.4 hours is now bought off the system by high "scarcity" prices. This represents the realizations of "State 2" conditions displayed in Figure 2. Demand response effectively flattens the very top of the load duration curve. This also leads to a change in the short-run marginal cost and distribution of the hours when each technology is marginal. See Table 7. There are now fewer hours when peaking capacity (MC = 80) is marginal and as much as 20.4 hours when demand response is marginal (MC = 4000). As already noted, we refer to these 20.4 hours either as "scarcity hours" or "shortage hours." Table 8 recalculates revenues, costs, and any shortfall in cost recovery using the expanded set of short run marginal costs, associated prices, and load durations. The major difference between Table 8 and Table 4 is that all generating capacity now receives $4000/Mwh during about 20 hours of "scarcity" conditions. As indicated by Table 8, with "scarcity pricing" during only 20 hours in the year, each generating technology now covers its total costs as does the system as a whole.

The "scarcity price" of $4000/Mwh may seem like either a lot to pay for avoiding reducing electricity consumption or (equivalently) too small a number of hours of the year for the system to be in "scarcity" or "shortage" conditions. In this example, if we reduced the value of lost load to $2500/Mwh, demand response would be triggered for about 33 hours and the maximum quantity of demand response would be 45 Mw, a qualitatively similar result.[11] It is important to recognize that the VOLL in this case reflects a very short run demand elasticity and (typically) a loss of load with little or no notice to the retail consumer and lasting for a few hours. Measuring the value of lost load

[11] If I had drawn the load duration curve in the example to have a higher "needle peak" demand lasting a 20-30 hours, the quantity of demand response would, of course, be larger. The cutoff operating duration would not change, however.

empirically absent meaningful market valuation data is a very difficult exercise. The VOLL will depend on the nature of the consumer activities interrupted, the notice that consumers are given before an interruption takes place (Joskow and Tirole (2005a)), whether the interruptions are voluntary through market arrangements or involuntary through rolling blacouts, and the duration of the outage.

Nevertheless, there have been numerous efforts to measure the value of lost load. Bushnell (2005, page 14) points to a range of estimates between $2,000 and $50,000/Mwh. Cramton and Stoft (2006, p. 33) suggest that conventional "planning" reliability criteria based on keeping the probability of rolling blackouts very low imply a value of lost load of $267,000/Mwh. The number of "scarcity" or "shortage" hours derived in the example presented here are also similar to those experienced in practice (Cramton and Stoft (2006, p.40) Accordingly, the numbers used in this numerical example are well within the range of available estimates from customer surveys and those implied by historical electricity system behavior and probably on the low side.

Clearly, the availability of demand response (demand elasticity) allows supply and demand to be balanced at a price that reflects consumers' willingness to pay for more or less supply in the very short run and satisfies a break-even constraint necessary to attract investment consistent with a least cost generation investment and operating equilibrium. For future reference, note that the revenues earned under scarcity conditions from "scarcity pricing" of energy in this example represent a large fraction of the quasi-rents necessary to cover the capital costs of the least cost quantity and mix of generating capacity. Table 9 displays the fraction of the quasi-rents earned from market revenues under "State 1" short-run marginal cost pricing conditions and "State 2" under scarcity

conditions. Both sources of rents are required to cover the capital costs of all three supply technologies that make up the least cost supply portfolio. For base load technologies 33% of the rents come from scarcity pricing, for intermediate load technology 50%, and from peaking technology 100%.

The failure to include active price-related demand response in this way or to keep prices from rising to $4000, for example by imposing price caps, 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. 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 define some default price or price cap at which suppliers will be compensated for energy and operating reserves under these conditions. This is the case because absent the availability of demand response to clear the market when demand reaches 22,000 Mw there is a vertical demand and vertical supply curve and there will by no well-defined market clearing price. Investment will adapt to whatever default pricing arrangements are chosen in this case. Assume that the system operator can implement a non-price rational scheme (i.e. rolling-blackouts) when capacity constraints are reached (rolling blackouts) to balance demand with the capacity constraint and sets the 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 example with demand response. 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.

While these numerical examples are static, the presence of uncertainty does not change the basic economics of investment and operation discussed above. Investment decisions would in principle reflect the expected values of the relevant variables on the demand and supply sides, including any risk bearing costs borne by consumers and/or investors. The value of lost load would be reflected in both investment and operating decisions. Uncertainty will also introduce volatility into both prices and profitability (quasi-rents) realized in spot energy and operating reserve markets. When peak demand is at the high end of the probability distribution peak period prices and profits will be relatively high and vice versa. However, in a least cost equilibrium the expected net revenues earned over time during "scarcity" conditions should still be equal to the carrying costs of a peaker, and the similar quasi rent results for the other technologies will also hold over time. As I will discuss, however, price formation during scarcity conditions in the presence of operating reserves, related reliability constraints and discretionary behavior by system operators can complicate significantly the market price formation process and the production of quasi-rents consistent with a least cost investment portfolio that meets administratively determined reliability criteria.

IS THERE AN INVESTMENT PROBLEM?

At first blush, some may find it surprising that policymakers are concerned that wholesale market mechanisms will not provide adequate incentives for investment in new generating capacity. The early experience with electricity sector liberalization during the

1990s suggested that competitive wholesale markets could and would mobilize adequate (or more than adequate) investment in new generating capacity. Substantial amounts of capital were mobilized during the late 1990s to support construction of new generating capacity in many countries that had implemented reforms. In the U.S., over 230,000 Mw of new generating capacity went into service between 1997 and 2005, most of it merchant capacity burning natural gas, an increase of 30% from the stock of generating capacity that existed in the U.S. in 1996. The net summer capability of generating capacity in the U.S., increased over 25% between 1997 and 2005 after taking account of both new entry and retirements (See Table 10). About 40% of the stock of generating plants in service in England and Wales at the time its electricity sector was restructured was replaced with modern efficient combined-cycle gas turbine (CCGT) technology between 1990 and 2002 as old mostly coal-burning generators were retired and replaced by what was expected to be less costly CCGT capacity. Many other countries implementing reforms during the 1990s, including Argentina, Chile and Australia, also attracted significant investment in new generating capacity (Jamasb 2002) after the reforms were initiated.

So, why are policymakers so concerned now? First, we should recognize that liberalization has evolved in much of Europe during a period when there was significant excess generating capacity, Spain and Italy being the major exceptions. Capacity constraints have not been on the policymakers' radar screen until recently. Even in England and Wales, the quantity of generating capacity in service today is not much greater than it was in 1990, with most of the investment in generating capacity during the 1990s being stimulated by opportunities to replace the inefficient stock of old generators operated by the state-owned CEGB at the time of privatization, expectations that natural

gas prices would stay low, long term contracts entered into by retail suppliers early in the UK's liberalization program, and the high prices for energy and capacity payments available in the wholesale market, inflated by the exercise of market power by the dominant generators (Wolfram). These investments were not the result of a significant need for new generating capacity to meet rapidly growing peak demand.

Second, the environment for financing new generating capacity has changed dramatically in the last few years as a result of financial problems faced by merchant trading and generating companies in Europe, the U.S., Asia and Latin America (Joskow (2005), Jamasb (2002), De Araujo (2001), Sioshansi and Pfaffenberger (2006)). Potential investors have gone to great lengths to convince policymakers that they will not provide investment funds for merchant generating capacity in the future under traditional project financing arrangements without major changes in the behavior and performance of wholesale markets. Whether they are crying wolf or signaling the reality of investor views, their arguments have increased policymakers' concerns about "resource adequacy" or "supply security."

Most importantly, as demand has grown, as older plants retire, and as wholesale market prices have risen, policymakers in many countries see little evidence of a response to these market signals in the form of investment in new merchant generating capacity. The situation in the U.S. has attracted particular concern by policymakers in those areas of the country where the electricity sectors have been liberalized and rely on merchant investment. After peaking at 55,000 Mw of new capacity entering service in the U.S. in 2002, the quantity of new generating capacity entering service in the U.S. and the quantity under construction has steadily declined. In 2005, only 15,000 Mw of new

generating capacity entered service, most of which was built either by municipal utilities that have not been subject to restructuring and competition reforms, by traditional vertically integrated utilities in states that have not liberalized their electricity sectors or wind projects that benefit from special subsidies and contractual arrangements. Concerns about investment in additional generating capacity to meet growing demand have been raised in New England, New York, PJM, and California. System operators in the Northeastern U.S. and California are projecting shortages and increases in power supply emergencies starting in two to three years, recognizing that since developing, permitting and completing new generating plants takes several years if there is little under construction today little will come out of the pipeline two or three years from now.

On the one hand, a market response that leads prices (adjusted for fuel costs) and profits to fall and investment to decline dramatically when there is excess capacity, is just the response that we would be looking for from a competitive market. At least some of the noise about investment incentives is coming from owners of existing merchant generating plants who would just like to see higher prices and profits. On the other hand, numerous analyses of the performance of organized energy-only wholesale markets in the U.S. indicate that they do not appear to produce enough net revenues to support investment in new generating capacity in the right places and consistent with the administrative reliability criteria relied upon by system operators and regulators.

The theoretical framework and the numerical examples in the last section make it clear that in order to attract investment to balance supply and demand with traditional levels of reliability, competitive wholesale markets must produce “rents” over and above the short-run marginal cost of operating generating facilities in order to provide

compensation for the capital costs of these facilities. Prices and the associated revenues produced during "scarcity" conditions when generating capacity is fully utilized are especially important. In particular, over time, wholesale prices must produce rents greater than or equal to the capital costs associated with marginal investments in new peaking capacity consistent with the least cost quantity and mix of generating capacity to balance supply and demand. Accordingly, a common test for whether wholesale markets are providing adequate price signals is to calculate the net revenues (quasi-rents) that would have been earned by a hypothetical investment in new peaking capacity from economical sales of energy and operating reserves over a period of several years.

The experience in the PJM Regional Transmission Organization (RTO) in the U.S. is fairly typical. Table 11 displays the net revenue that a hypothetical new combustion turbine would have earned from wholesale energy market plus ancillary services revenues in PJM if it were dispatched optimally to reflect its marginal running costs in each year 1999-2005. In no year would a new peaking turbine have earned enough net revenues from sales of energy and ancillary services to cover the annualized capital costs of a new generating unit and, on average, the net revenues contributed only about 40% of the annualized capital costs of a new peaking unit.[12] Based on energy market revenues alone, it would not be rational for an investor to invest in new combustion turbine capacity in PJM based on six years of historical experience. Similar calculations of net energy market revenues have been performed for hypothetical investments in new CCGT capacity and pulverized coal capacity in PJM. These calculations also indicate that energy market revenues alone do not come close to

[12] These calculations are probably an overestimate of the net revenues that a new peaking unit would realize in practice since that assume "perfect" economic dispatch and do not take account of various operating constraints (PJM 2006, pp. 128-132).

covering the capital costs of new investments in these technologies either (PJM 2006, 127-132). This net revenue gap or the "missing money" problem referred to by Cramton and Stoft (2006) is a major deterrent to investment in new generating capacity in the organized wholesale markets in the U.S. today.

As I will discuss on more detail in the next section, one solution to the "missing money" problem in the U.S. has been to impose capacity obligations on load serving entities,[13] to create a market for the associated qualifying capacity, and in this way to create another stream of revenues for generators that it has been hoped would make up for the net revenue gap in the energy market. For example, load serving entities (LSEs) might be required to have contracts for qualifying generating capacity equal to 118% of their peak load each year. The 18% reflects a capacity reserve margin defined to meet reliability criteria established by the reliability authorities in the area in which the LSEs purchase power. There is then a market for qualifying capacity that defines capacity prices. Indeed, PJM has always had capacity obligations which it carried over into its competitive market design.

In theory, capacity prices should adjust to clear the market consistent with the reserve margin chosen and make up for the "missing money" (Joskow and Tirole (2005a)). However, even adding in capacity-related revenues in PJM during the six year period covered by Table 11, the total net revenues (energy plus capacity related revenues) that would have been earned by a new peaking unit over this six year period were significantly less than the capital costs that investors would need to expect to recover to make investment in new generating capacity profitable. The average annual capacity

[13] Load serving entities include distribution companies with retail supply obligations and competitive retail electricity suppliers.

market revenue for a combustion turbine in PJM from 1999-2005 was about $13,000/Mw/year (PJM (2006), pp. 230-232). Adding the capacity market revenues to the net revenues from sales of energy and ancillary services in Table 11 brings the total net revenues for a hypothetical peaking unit to about $40,000/Mw/Year for 1999-2005, roughly $35,000/Mw/Year short of the annualized capital costs of new peaking capacity. Again, similar results are revealed for CCGT and pulverized coal technology investments.

This "missing money" phenomenon is not unique to PJM. Every organized market in the U.S. exhibits a similar gap between net revenues produced by energy markets and the capital costs of investing in new capacity measured over several years time (FERC (2005), p. 60; New York ISO (2005), pages 22-25, Joskow 2005). Indeed, since 1998 there isn't a single year when energy market revenues covered the annualized capital costs of a peaking turbine. There is still a significant gap when capacity payments are included. The only exception to the latter result appears to be New York City where prices for energy and capacity collectively appear to be sufficient to support new investment, though new investment in New York may be much more costly than assumed in these analyses (FERC (2005), page 60). Moreover, a large fraction of the net revenue estimated for investment in generating capacity in New York City comes from capacity payments rather than energy market revenues (New York ISO (2005), p. 23).

One potential explanation of these results is that they simply imply that there is excess generating capacity in these systems and the low net revenue results are simply signaling that two much capacity has come into service. That is, this is an indicator of excess generating capacity. However, this result is inconsistent with the behavior of

system operators in the Northeaster U.S., California, and in other countries which are forecasting capacity shortages in the near term and are taking actions to stimulate more investment. For example, the New England ISO forecasts significant capacity needs beginning in 2008, but there is almost no new generating capacity under construction at the present time. Moreover, in New England the energy and capacity markets are not even producing enough net revenues to keep a significant amount of generating capacity from closing down (typically permanently). The New England ISO has found it necessary to sign special "reliability contracts" for up to 7,000 Mw of existing generating capacity to keep it in service (ISO New England (2005), page 80). PJM also forecasts that there will be a need for a significant quantity of new generating capacity to meet demand in the next few years. The generating capacity now under construction does not satisfy these forecast needs, which are magnified by plans by old generating units to retire. Thus, the failure of wholesale markets to provide adequate revenues is the primary suspect for the failure of investors to begin to build new generating facilities to match forecasts of resource needs.

A more subtle counter-argument is that policymakers are overestimating the need for additional generating capacity because these estimates are based on old reliability criteria that do not properly reflect consumer valuations. That is, the reliability criteria used by the reliability organizations in the U.S. (and other countries since they are very similar) are inconsistent with the marginal value of lost load to consumers during these periods. According to this view, the market is signaling that consumers do not want to pay for this much reliability and the market, rather than reliability organizations, should make that choice. It may very well be that reliability targets require more generating

capacity than consumers are willing to pay for and that these engineering reliability criteria should be reevaluated. However, as I will discuss further below, at the present time it is unlikely that market mechanisms have yet evolved to produce the appropriate level of operating reserves or capacity margins consistent with consumer valuations of lost load resulting from potential rolling blackouts and network collapses. Moreover, reducing reliability in these dimensions is not politically appealing and, in the U.S., runs counter to the provisions of the Energy Policy Act of 2005 which seek to strengthen, harmonize, and enforce traditional reliability criteria more aggressively.

WHAT ARE THE CAUSES OF THE "MISSING MONEY" PROBLEM?

The ultimate source of the "missing money" problem is that spot market prices do not rise high enough during "scarcity" hours to produce adequate quasi-rents to cover the capital costs of investment in an efficient level and mix of generating capacity. Since prices for forward contracts reflect the expected value of spot market prices (plus any risk-bearing costs) via intertemporal arbitrage, any truncation of the upper tail of the distribution of spot prices will be reflected in forward prices that are below the efficient level as well. But why don't wholesale markets produce adequate revenues? There are a number of wholesale market imperfections, regulatory constraints on prices, as well as procedures utilized by 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 hours in a typical year when they should be very high.

To understand the sources of the missing money problem we must examine in more detail how system operators in the organized markets in the U.S. balance supply and demand on real electric power networks, especially during "scarcity" or "operating reserve shortage" hours. In a market context, the attributes of the price formation process during these operating reserve shortage conditions is critical for understanding whether and how the wholesale market provides appropriate price signals to attract investment. If it were the case that operating reserve constraints were always met by variations in prices that kept supply and demand in balance continuously, as in simple theoretical models of electric power systems with demand response, then there would be no problem. Indeed, there would be no need for system operators to establish operating reserve and other reliability criteria. The market could be relied upon to do so. However, at least at the present time, there are a number of market imperfections that make it unlikely that markets will lead to this happy result:

a. Only a tiny fraction of electricity consumers and electricity demand during peak hours can see real time prices and 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. As a result, on a typical U.S. network 98+% of peak demand is effectively price inelastic in the time frame that system operators are looking for during scarcity conditions. Since supply is also effectively up against capacity constraints during operating reserve deficiency conditions we face a situation where we have a vertical demand curve and a vertical supply curve. Under these conditions system operators in the U.S. resort to non-price rationing of

demand (rolling blackouts) to maintain minimum operating reserve levels and the frequency, voltage, stability and other physical engineering operating reliability criteria.

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 (2005b)). 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 (2005b)). 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. Zonal rationing is especially problematic in the presence of retail competition (Joskow and Tirole (2005a, 2005b)) and gives reliability as reflected in the probability and duration of demand curtailments collective good attributes.

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 a network collapse such as those that occurred in the Northeastern U.S. and in Italy in 2003 very low. When there is a network collapse there is both excess demand and excess supply because the network infrastructure to allow demand and supply to interact has collapsed. The outages are widespread and restoring the system to operational status can be time consuming and costly. Nevertheless, since the market also collapses in these situations prices are effectively zero. 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 (2005a) 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.

These three attributes of electric power networks give reliability public goods attributes. Accordingly, even if the other market, regulatory, and behavioral imperfections are resolved we cannot count on “the market” alone to provide the efficient level of reliability.

d. Rolling blackouts resulting from a shortage of generating capacity are extremely rare on electric power systems in developed countries.[14] 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

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

triggers non-price rationing of demand. The value for additional scarcity rents earned under scarcity conditions are uncertain since they depend on the operating protocols implemented by the system operator during operating reserve deficiencies and the associated price formation process.[15] 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 market mechanism exists to produce the efficient price levels during these hours. (Joskow and Tirole (2005a)). And a close examination of system operator protocols and behavior during scarcity conditions makes it fairly clear that it is highly unlikely that efficient "scarcity prices" will emerge during operating reserve shortage contingencies. I offer two examples here.

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 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

[15] 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).

market prices. As long as voltage reductions are employed in this way, market price signals will lead to underinvestment in reliability because the social costs of voltage reductions are not internalized.

Second, 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 market for spinning reserves may not have a locational dimension to it or it may reflect a very crude distinction between geographic zones. Generator attributes are typically much more differentiated within the general product definitions used in organized wholesale markets in the U.S. 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. The behavior of the New England ISO during a severe cold snap in January 2004 is an example of this behavior and its consequences (FERC (2005), p. and ISO New England (2004)). Despite the fact that the New England electric

power network was severely stressed during this period, prices did not rise to levels that produced market-based quasi-rents for either CCGTs or peaking turbines; the spark spreads were zero or negative.

e. The limited amount of real time demand response in the wholesale market leads to 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 (2005a)).

If there is a significant unmitigated market power problem then wholesale prices should be too high. But the analysis above suggests that wholesale prices are too low not too high on average. As a result, many economists assume that the primary source of the "missing money" problem must be the price caps and related market power mitigation procedures imposed by regulators. That is, that the efforts to mitigate market power have had the effect of suppressing energy prices too much, especially during scarcity conditions when prices should be very high.[16]

The problem with blaming the entire problem on the price caps is that when one examines the full distribution of energy prices in the organized U.S. wholesale energy and operating reserve markets over the last six year it is evident that the price caps, which do in fact appear too low compared to estimates of the value of lost load, 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 cap are the only source of the missing money problem. I believe that the effects (not the goal) of the other system operator behavioral factors discussed above 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.

There also exist de facto price caps on capacity prices in those wholesale markets in the U.S. that have implemented capacity obligations and associated capacity markets. The way these markets have worked historically, the penalty imposed on LSEs for not contracting for adequate capacity, has been a monthly or annual deficiency charge

[16] Price caps that constrain prices to levels below competitive market prices in some periods but allow prices to rise above competitive market levels in other periods do not necessarily lead to a shortage of generating capacity. In this case, however, price caps would induce the wrong mix of generating capacity.

assessed by the system operator. The deficiency charge is typically calculated based on the annualized lifetime capital cost of a new peaking turbine using a set of assumptions about the cost of capital, depreciation, plant life, and taxes. This approach appears to be consistent with the discussion of the quasi-rents that must be earned by a peaking turbine to make competitive entry financially attractive and to support least cost investment in all technology options. In practice, however, it is not because it assumes implicitly that capacity will earn net revenues equal to the deficiency charge in each year of its economic life.

The capacity obligations that are central to these systems have historically relied on hard reserve margin criteria (e.g. 18% of peak load). Due to uncertainty on both the demand and supply sides, even if the target reserve margin is hit on average over a period of years, there will be some years when the actual reserve margin is greater than the target and some years when it is less than the target. In those wholesale markets with capacity obligations, capacity prices have tended to rise to the level of the deficiency charge during periods when supplies are tight and then drop to zero or close to it during periods when the reserve margin exceeds the target. On average, the revenues are then significantly less than the lifetime carrying charges of a peaker. If the distribution of realized reserve margins is symmetrical around the target, generators will earn only 50% of the capital costs of a peaker over time. Thus, by calculating the deficiency charge in this way, a de facto price cap is placed on capacity prices as well. This is the primary reason why traditional capacity obligations and capacity markets have not solved the missing money problem.

OTHER POSSIBLE DETERRENTS TO GENERATION INVESTMENT

While in my view the "missing money" problem is the most serious deterrent to investment in generating capacity, other financial barriers to efficient investment in generating capacity have also been identified by various commentators. Wall Street investment bankers routinely argue that investment in new generating capacity will not be forthcoming because prices in wholesale spot markets are too volatile and there are inadequate opportunities for investors to find counterparties willing to enter into forward contracts of ten or more years duration to allow investors to hedge market risks. They claim that absent long term contracts with creditworthy buyers it will be difficult to find financing for any merchant generating project.

I don't know of any good theoretical reason why market price volatility or price uncertainty per se should make it impossible to finance new generating facilities if the "missing money" problem is solved. Perhaps price uncertainty will affect the cost of capital used by investors to evaluate projects, but this would just increase the prices and quasi-rents that the market would have to produce to stimulate investment. Investors finance oil refineries, oil and gas drilling platforms, cruise ships, and many other costly capital projects where there is considerable price uncertainty without the security of long term contracts.

One attribute of electricity markets that may have implications for the efficient allocation of market price risk between investors, intermediaries and consumers is the retail procurement framework that has accompanied the liberalization of wholesale electricity markets. In the U.S. and several other countries, comprehensive retail competition programs have been created but have been slow to evolve. Large fractions

of system demand continue to be served by incumbent distributors with default service obligations and who contract for power with relatively short-term contracts. The contracting requirements are driven by regulatory requirements rather than through market-based allocations of risk. There is no reason to believe that they are optimal. As retail competition matures and retail suppliers with large diversified portfolios emerge, they are likely to be more willing voluntarily to take on longer term commitments to buy power from generators (or build their own generating portfolios) if this can reduce the prices they must pay to buy power over time. While individual retail consumers may only have one, two or three year contracts, a diversified portfolio of retail customers, especially smaller customers who are reasonably "sticky," would provide a retail supplier with the kind of stable demand base that it would need to make it potentially attractive to sign long term supply contracts.

This observation leads directly to questions about the optimal contractual, financing market structure for electricity suppliers at wholesale and retail. The initial model for independent power producers that emerged in the U.S. after the Public Utility Regulatory Policy Act (PURPA) went into effect in the early 1980s, was based on long-term purchase contracts between independent power producers and regulated utilities. Project financing with high debt/equity ratios secured by these contracts was the financing framework of choice. The next wave of investment in merchant generating capacity beginning in the late 1990s relied on the project financing model but without the long term contracts. When wholesale markets collapsed after 2001 many of those projects could not meet their debt obligations and many went bankrupt or were subject to alternative financial restructurings.

I believe that the merchant investment model based on wholesale generating companies relying on highly leveraged individual project financing arrangements is likely to be poorly suited to a competitive wholesale and retail market framework. Partial vertical integration between retail supply and generation ownership (but not T&D) combined with diversified portfolios of spot, short and medium term contracts with independent suppliers to make up for the rest of the retail supplier's wholesale power requirements, is likely to be a superior organizational form for financing investment and dealing with imperfections in wholesale spot markets, including the potential "hold-up" problems that I will discuss presently. Such vertically integrated retail supply and generation companies are likely to be large firms with substantial balance sheets and rely primarily on balance sheet financing for their generation portfolios. The power supply industry will look more like the oil and gas industry, with a relatively small number of large vertically integrated firms, and a large number of "small" independent generating companies. This industrial structure is gradually emerging in the U.S. and Europe. There need not be a conflict between competition goals and an industry with large vertically integrated power supplies as long as the firms' wholesale and retail supply businesses are sufficiently dispersed geographically that there are several competing suppliers in any region and the transmission network is owned and operated independently. However, in practice there may be a conflict between vertical integration and competition in regional markets where there is a dominant vertically integrated incumbent and associated barriers to entry of competing vertically integrated suppliers.

A final reason why it may be difficult to finance investments in new generating capacity are concerns about opportunistic behavior by government regulators or system

operators that may affect spot market prices at critical times over the life of a new generating unit. As discussed in detail above, a large fraction of the net revenues or quasi-rents from sales of energy in spot electricity markets required to cover the costs of capital investments is produced in a very small number of hours each year when capacity is fully utilized. Moreover, due to uncertainty on the demand and supply sides, these hours will not appear uniformly from year to year but will fluctuate widely from year to year. One year it may be 80 hours and another year 5 hours of scarcity conditions (Joskow (2005), Cramton and Stoft (2006, p. 33)). For a peaking plant, all of its net revenues are derived under these conditions. Accordingly, investors must be very concerned about actions by regulators or discretionary behavior by system operators that might have the effect of constraining prices in exactly those few hours with very high prices when investors expect to earn most of the net revenues required to cover their capital investment costs. It is now widely recognized that opportunism problems, whether by counterparties or government entities, can lead to under-investment and that credible long-term contracts or vertical integration are efficient institutional responses to opportunism problems (Williamson (1979), Hart (1985), Joskow (1987)). From the investor's perspective, long term power supply contracts with credit worthy buyers can allow them to shift this risk to buyers.

POLICY RESPONSES

Numerous policy proposals have been made to fix what is now widely viewed in the U.S. as the failure of organized wholesale power markets to provide adequate incentives to stimulate investment in new generating capacity to balance supply and

demand efficiently consistent with system reliability criteria. I will focus primarily on the missing money or net revenue gap problem here. However, I will also take into account related concerns about market power mitigation, price volatility, and opportunism. The proposed policy reforms involve a combination of mechanisms to improve the performance of spot markets so that prices will come closer to reflecting the (uncertain) value of lost load during scarcity conditions and a forward market for reserves that reflects the reliability targets specified by regulators. These reforms involve price triggers and quantity targets and are related to the application of a combination of prices and quantities to controlling pollution discussed by Roberts and Spence (1976).

a. Improving the performance of organized spot markets: The fundamental source of the net revenue gap problem is the failure of spot energy and operating reserve markets to perform in practice the way they are supposed to perform in theory. It is natural to focus on improving the performance of these markets. 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 market's provision of reliability, 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.

i. Raise the price caps and hit them during scarcity conditions:: The $1000/Mwh price cap in effect in most of the organized markets in the U.S. ($250/Mwh in California) is a completely arbitrary number that is clearly below what the competitive market clearing price would be under most scarcity conditions (State 2). However, as I

have discussed, the $1000 price caps are rarely binding constraints in the organized U.S. markets so that increasing them alone would not have much of an impact on the net revenue gap problem. Increasing the price caps to reflect reasonable estimates of VOLL would also make it more attractive and profitable for suppliers to exercise market power in spot energy and operating reserve markets. Nevertheless, there are good reasons to increase the price caps to reflect reasonable values of VOLL if this is combined with changes to the wholesale market 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.

To make the higher price caps meaningful contributors to the net revenue gap problem and to deal with the price formation problems that emerge when system operators implement reliability protocols when there are capacity constraints, I would propose that whenever a system operator issues a notice that operating reserve deficiency protocols will be implemented 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 to where they should be under scarcity conditions and is a practical response to the challenges of integrating reliability rules, responses like voltage reductions which are not properly priced through market mechanisms, and various discretionary behavior that we must allow system operators to undertake to maintain network reliability and avoid network collapses.

As with raising the price caps, this increases supplier incentives to withhold supplies as capacity constraints are being approached and market monitors will have to focus their attention on withholding of capacity during hours when capacity constraints

are being approached. However, there are mechanisms other than price caps that can help to mitigate market power. It has been widely recognized that more reliance on forward contracting for energy can help to mitigate spot energy market power problems (Wolak (2004), Allaz and Vila (1993)) and there have been many recommendations that wholesale markets should rely much more on forward contracting. More forward contracting would be a good thing from both a market power mitigation perspective and from the perspective of those who believe that price volatility, price uncertainty, and opportunism are deterrents to investment.

One problem here is that proponents of more forward contracting provide little guidance regarding how this goal will be achieved in the context of retail competition. With competitive retail markets it is generally up to retail customers and their supply intermediaries to decide on their contractual arrangements, including contract duration. If retail suppliers are not voluntarily entering into longer term contracts we need to understand why and if implementing the recommendation that more reliance be placed on long term contracts involves compelling LSEs to enter into bilateral forward contracts with generators, the implications of doing so also need to be better understood; in particular the implications for the diffusion of retail competition. I will discuss below how the creation of a forward capacity obligation and associated capacity markets can be structured to also hedge energy prices during peak periods and mitigate incentives to exercise market power.

ii. 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[17] can also help both to increase the efficiency with which capacity constraints are managed and improve the price formation process during scarcity conditions. However, the way in which demand response is brought into the system for these purposes is important. Demand response should be integrated into the system in a way that is symmetrical to the treatment of supplies of energy, operating reserves, and capacity. 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. If capacity payments are made to generators then equivalent capacity payments should be made to qualifying demand response. It also matters exactly how capacity payments are reflected in retail prices (Joskow and Tirole (2005a)). Today, demand response resources tend to be pre-contracted, the costs partially recovered through uplift charges spread over many hours, and calls on demand response triggered by system operating conditions and reliability protocols rather than high prices. The New York ISO has done a good job improving the ways in which demand response is integrated into spot energy markets and this is the kind of reform that I have in mind (New York ISO (2005b)).

iii. 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

[17] 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.

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 in the bid-based supply stack. The supply of energy and various operating reserve services are substitutes, arbitrage links their market prices together, and opportunities exist to change the use and physical attributes of generating facilities in response to price incentives for specific operating reserve attributes.

iv. 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 part efforts 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. The criteria used for operating reserve targets may also be inconsistent with consumer valuations. 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.

b. Capacity obligations, forward capacity markets and capacity prices[18] The reforms to wholesale energy markets discussed above should help to reduce the net revenue gap. However, it is not at all obvious that the missing money problem will be 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 LSEs and the effective design of associated capacity markets. If properly designed, forward capacity markets can act as a safety valve to fill the net revenue gap and support efficient investment in generation and demand response, are compatible with the continued evolution of wholesale spot markets, are consistent with the continued evolution of retail competition, and can help to reduce investor concerns about price volatility and opportunism. If spot energy and ancillary reserve market performance improves dramatically, capacity obligations and capacity markets can also effectively fade away.

i. forward contracts for energy alone do not solve the net revenue gap or missing money problem: Before discussing how forward capacity obligations and associated capacity markets can be structured to do all of these good things I want to briefly discuss one type of frequently mentioned proposal that will not in and of itself solve the net revenue gap problem. Several proposals have been made to require LSEs (or system

[18] The new forward capacity market framework filed in March 2006 with FERC by the New England ISO as a settlement among many parties contains many of these features (ISO New England (2006)) . See Cramton and Stoft (2006) for a detailed discussion of the rationale for the provisions of the New England ISO's forward capacity market proposal.

operators) to enter into some type of "hedged" forward contracts for energy to cover a large fraction of their retail customers' energy demand. The proposals include fixed price forward contracts for energy between LSEs and generators as well as option contracts that specify a call price for energy ex ante (e.g. Wolak (2004), Oren (2005)). It is claimed that these hedging contracts will solve the "resource adequacy" problem. These assertions are simply wrong.[19] They are wrong because they do not deal with the underlying market imperfections and institutional constraints that lead to the missing money problem and implicitly assume, without explanation, that the relevant market failure results from inadequate forward contracting by retail consumers and their retail suppliers. They ignore the considerations discussed above that lead to the conclusions that "the market" cannot be relied upon to select the optimal level of reliability.[20] Moreover, policymakers will not allow the market to make this choice. They will continue to impose reliability standards and associated operating reserve requirements and capacity reserve requirement criteria as they do now. There may be good reasons to change these requirements and the mechanisms utilized to meet them, but economists are dreaming if they think that policymakers will be ready soon to leave reliability criteria to the market. The hedging contract proposals do reduce price volatility and are likely to mitigate market power. These are good outcomes. However, unless they incorporate generating capacity reserve criteria ("resource adequacy criteria") as well they will not solve the missing money problem. The forward contract prices will just reflect the low spot wholesale energy prices that create the net revenue gap in the first place.

---

[19] Bidwell (2005) and Singh (2000) also have made proposals that have option contract components. However, they also have components that deal with the missing money problem by incorporating reliability criteria. See Cramton and Stoft (2006) for a more detailed comparison of these proposals.

[20] See Cramton and Stoft (2006) which discuss this issue in more detail, focusing on the implications on limited real time pricing and the inability to control power flows to individual consumers.

ii. Implement well-designed forward capacity markets: Recent so-called "capacity market" proposals start with the reliability criteria established by the responsible reliability organizations.[21] The primary generating capacity-related criterion is typically a generating capacity reserve margin measured by the difference between the system peak demand (D) before any curtailments and the peak generating capability (G) of 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 capability 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. 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. 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 peaks. All 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. Under the forward capacity market proposal the auctions are for delivery several years into the future and prices may be fixed, at the supplier's choice, for a few years starting with the delivery date.

LSEs can meet their forward capacity obligations either by contracting directly with generators for capacity to be available to supply energy at the time of system peak or

[21] 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.

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"). 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. Owners of generating capacity that clears in the market has 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.

Under the forward capacity market proposal the spot energy markets continue to operate as before, with whatever improvements are introduced as discussed above. Following the numerical examples above, in equilibrium the market clearing price (P_c) for generating capacity should equal the capital costs of a peaker (P_k) less the quasi-rents that a peaker would expect to earn (R_p) in the energy market or $P_c = (P_k - R_p)$ adjusted for

expected forced outage rates and associated penalties (Joskow and Tirole (2005a)).[22] This solves the missing money problem since the capacity price essentially acts as a safety valve to fill the gap between the capital costs of a peaker and the quasi-rents that a peaker expects to earn in the energy and operating reserve markets. Moreover, as the performance of the wholesale spot energy market improves, the expected quasi-rents produced for a peaker in the energy market will rise toward $R_p = P_k$ and the capacity price will fall toward zero.

As already noted, simple versions of capacity obligation/capacity market approach have been operating for years in several U.S. ISOs, but have not solved the missing money problem or the other problems noted above. The forward capacity market proposals on the table today include several enhancements to these older capacity mechanisms. I now discuss several of the enhancements that characterize the forward capacity market framework.

The earlier 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 from wholesale energy and operating reserve markets. The enhanced mechanisms 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, rather than being set arbitrarily at the average annualized lifetime capital costs of a peaker, so that on average the mechanism should yield a capacity price equal to P_k before netting out any quasi-rents produced in the energy market. The proposed annual capacity price cap included in the forward

[22] 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.

capacity market proposal for New England is more than twice the old deficiency payment cap.

A second problem noted with the existing 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 zero and the price cap from year to year. The New York ISO has introduced a reserve capacity demand curve mechanism 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. A similar approach was proposed for New England. However, the initial proposal was renegotiated and, among other changes, the demand curve was replaced with an auction mechanism with caps and floors (a "price collar") on capacity prices and an intertemporal adjustment mechanism to reflect information about capacity market values drawn from actual market behavior over time. Together, these provisions also have the effect of smoothing out the distribution of capacity prices and better reflecting the value of capacity above and below a hard installed generating capability target.

A third problem sometimes identified with the existing capacity obligation/market arrangements is 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 reforms proposed 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 the New England proposal, the capacity auction will be for capacity that is to be available to the market over three years in the future.

A fourth problem identified with the existing 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 opportunism problems is unclear. However, the New England forward capacity market proposal allows new entrants at their choice to lock in capacity prices determined in the auction for a period of up to five years after the forward capacity delivery date.

A fifth problem identified with the existing 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 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 proposals include penalties for generators who are not available to perform when they are most needed.

A sixth problem identified with the existing 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.)

A final criticism of the existing capacity market arrangements is that they fail 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 New England proposal has 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 a hypothetical peaking unit with a high marginal generating 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

of capacity decreases. Second, it provides a net revenue hedge to peaking capacity that performs as expected and a partial hedge to base load and intermediate capacity. 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. As Cramton and Stoft (2006) argue persuasively, by hedging prices paid by load during peak hours this additional component of a forward capacity market design effectively integrates the forward contract/options/load hedging proposals discussed above within a framework that also deals with the missing money problem.

Most of the discussion of capacity obligation/market mechanisms has focused on the supply side. To fully restore appropriate incentives to market participants, the demand side of the market should be treated symmetrically. 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 retail consumers during these same critical periods. This should be a goal of further refinements in the forward capacity market framework.

Much of the discussion of proposals for dealing with generation investment incentives has also ignored the implications for the further evolution of retail competition. The proposals that would require LESs to enter into a portfolio of long term

contracts with individual generators for supplies of energy to meet their peak loads are in my view incompatible with retail competition. In areas where a large fraction of the retail load has not switched to competitive suppliers the responsible LSE would be the incumbent regulated distribution company. The costs of the long term contracts signed by the LSE would then be passed through to "default service" retail customers on a cost of service basis. This raises potential stranded cost problems (again) and can distort decisions by consumers regarding switching to competitive retail suppliers or default service as wholesale market prices will inevitably deviate from the average cost of the of the regulated incumbent's portfolio of long term contracts at any point in time that is used to set regulated default retail prices This approach also places additional financial burdens on competitive retailers since it will increase their credit obligations to become counterparties to long-term supply contracts. Retailers may not be able to put together a retail contract portfolio that matches their wholesale contract obligations or to recover the market value of the contractual risks that have been imposed upon them in market-based retail prices. Accordingly, requiring all LSEs to enter into long term contracts will increase the market risk faced by competitive retail suppliers placing an additional burden on the already slow diffusion of retail competition.

The forward capacity market mechanism is much more compatible with retail competition than are the proposals that place forward contracting proposals on individual LSEs. Capacity prices are set through an organized market process and the associated financial obligations to make the capacity payments are ultimately a collective obligation of all retail suppliers in the aggregate rather than a long term capacity commitment of each individual retail suppliers. As retail customers switch from retailer to retailer, the

capacity obligations associated with their demand move along with them along with the financial obligations (capacity prices) associated with the forward price obligations determined through forward capacity auctions. Individual retail suppliers do not have to post credit to support (say) five-year contractual commitments since the credit is provided by the collective obligations of retail suppliers defined in the system operator's tariff.[23] Since the obligations for capacity payments are based on each retail supplier's share of peak demand and the price of capacity is established ex ante though an auction mechanism, movements of retail customers among retail suppliers and the associated movement in capacity payment obligations can be handled easily by the system operator.

CONCLUSION

Evidence from the U.S. and some other countries indicates that organized wholesale markets for electrical energy and operating reserves do not provide adequate incentives to stimulate the proper quantity or mix of generating capacity consistent with mandatory reliability criteria. Based on U.S. experience, a large part of the problem can be associated with the failure of wholesale spot markets for energy and operating reserves to produce prices for energy during periods when capacity is constrained that are high enough to support investment in an efficient (least cost) mix of generating capacity. A joint program of reforms applied to wholesale energy markets, the introduction of well-design forward capacity markets, and symmetrical treatment of demand response and generating capacity resources is proposed to solve this problem. This policy reform program is compatible with improving the efficiency of spot wholesale markets, the continued evolution of competitive retail markets, and restores incentives for efficient

[23] All retail suppliers and generators would still have to meet the system operator's standard credit requirements and provisions for obligations incurred by suppliers who go bankrupt must be defined.

investment in generating capacity consistent with operating reliability criteria applied by system operators. This reform package also responds to investment disincentives that have been associated with volatility in wholesale energy prices by hedging energy prices during peak periods as well as responding partially 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.

REFERENCES

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

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.

Bidwell, Miles (2005), "Reliability Options," *Electricity Journal*, June, 11-25.

Bushnell, James (2005), "Electricity Resource Adequacy: Matching Policies with Goals," CSEM Working Paper No. 146, August.
http://www.ucei.berkeley.edu/PDF/csemwp146.pdf

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.

Hart, Oliver. 1995. *Firms Contracts and Financial Structure*. Oxford: Clarendon Press.

ISO New England (2004). "Final Report on Electricity Supply Conditions in New England During the January 14-16, 2004 Cold Snap." October. http://www.iso-ne.com.

ISO New England (2005). *2004 Annual Market Report.* http://www.iso-ne.com

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, Paul L. 1987. "Contract Duration and Relationship Specific Investments," *American Economic Review*, 77:168-75.

Joskow, P.L. (2005a). "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. and J. Tirole (2005a). "Reliability and Competitive Electricity Markets." September, (revised).
http://econ-www.mit.edu/faculty/download_pdf.php?id=917

Joskow, P.L. and J. Tirole (2005b), "Retail Electricity Competition," *Rand Journal of Economics* (forthcoming).

New York ISO (2005a), "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

New York ISO (2005b), "NYISO's Demand Respone Program," presentation by Aaron Breidenbaugh, March 15.

PJM Interconnection (2005), *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.

Oren, Shmuel (2005), "Generation Adequacy Via Call Options: Safe Passage to the Promised Land," University of California Energy Institute Working Paper EPE-016, September. http://www.ucei.berkeley.edu/pwrpubs/epe016.html and
http://stoft.com/metaPage/lib/Oren-2005-09-call-options-obligations.pdf.

Roberts, March and Michael Spence (1976), "Effluent Charges and Licenses Under Uncertainty," *Journal of Public Economics*, 5:193-208.

Singh, Harry (2000), "Call Options for Energy: A Market-based alternative to ICAP," mimeo, October. http://stoft.com/metaPage/lib/Singh-2000-10-Options-ICAP.pdf

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.

U.S. Federal Energy Regulatory Commission (FERC 2005). *State of the Markets Report*. http://www.ferc.gov/EventCalendar/Files/20050615093455-06-15-05-som2004.pdf

Williamson, Oliver (1979). "Transaction-cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, 22:3-61.

Wolak, Frank (2004), "What's Wrong with Capacity Markets," mimeo. http://stoft.com/metaPage/lib/WolaK-2004-06-contract-adequacy.pdf

Wolfram, Catherine (1999). "Measuring Duopoly Power in the Deregulated UK Electricity Market," *American Economic Review*, 89: 805-826.

TABLE 1

HYPOTHETICAL ELECTRIC GENERATION TECHNOLOGY OPTIONS AND LOAD DURATION CURVE

Generation Technology	Annualized Capital Costs $/Mw/Year	Operating Costs $/MWH
Base load	$240,000	$20
Intermediate	$160,000	$35
Peaking	$ 80,000	$80

Load Duration Curve (See Figure 1)

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

D = System load

H = Number of hours system load reaches a level D

TABLE 2

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

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	2,435	1 – 1778	$0.366
TOTAL	22,000		$7.694

TABLE 3

SHORT-RUN MARGINAL COST PRICING
PRICE DURATION SCHEDULE

Marginal Technology	Short-run Marginal Cost $/Mwh	Duration hours
Base load	$20	3427
Intermediate	$35	3556
Peaking	$80	1778

TABLE 4

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

Generating Technology	Revenues ($billions)	Total Cost ($billons)	Net Revenue Shortfall $(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
	$5.934	$7.694	($1.760)	

TABLE 5

HYPOTHETICAL ELECTRIC GENERATION SYSTEM WITH DEMAND RESPONSE "TECHNOLOGY"

Generation Technology	Annualized Capital Costs $/Mw/Year	Operating Costs $/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

TABLE 6

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

TABLE 7

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

TABLE 8

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

Generating Technology	Revenues ($billions)	Total Cost ($billons)	Shortfall $(billions)	Shortfall $/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-

TABLE 9

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

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

TABLE 10

U.S. GENERATING CAPACITY ADDITIONS
1997 – 2005

Year	New Generating Capacity (MW)
1997	4,000
1998	6,500
1999	10,500
2000	23,500
2001	48,000
2002	55,000
2003	50,000
2004	20,000
2005	15,000
	230,000

Total U.S. generating Capacity (MW net summer capacity):

1996	776,000
2005	980,000

Source: U.S. Energy Information Administration

TABLE 11

NET ENERGY AND ANCILLARY SERVICES REVENUES
NEW COMBUSTION TURBINE PEAKING PLANT
IN PJM

1999-2005

Year	Simulated Net Energy and AS Revenue $/Mw/Year
1999	$64,313
2000	18,724
2001	41,517
2002	25,480
2003	14,402
2004	10,311
2005	17,989
Average	$27,534

Annualized 20-year fixed cost ~ $70,000 - $80,000/Mw/Year

Source: *2005 State of the Market Report*, pages 124-132, PJM Interconnection

FIGURE 1

SHORT RUN MARGINAL COST PRICING

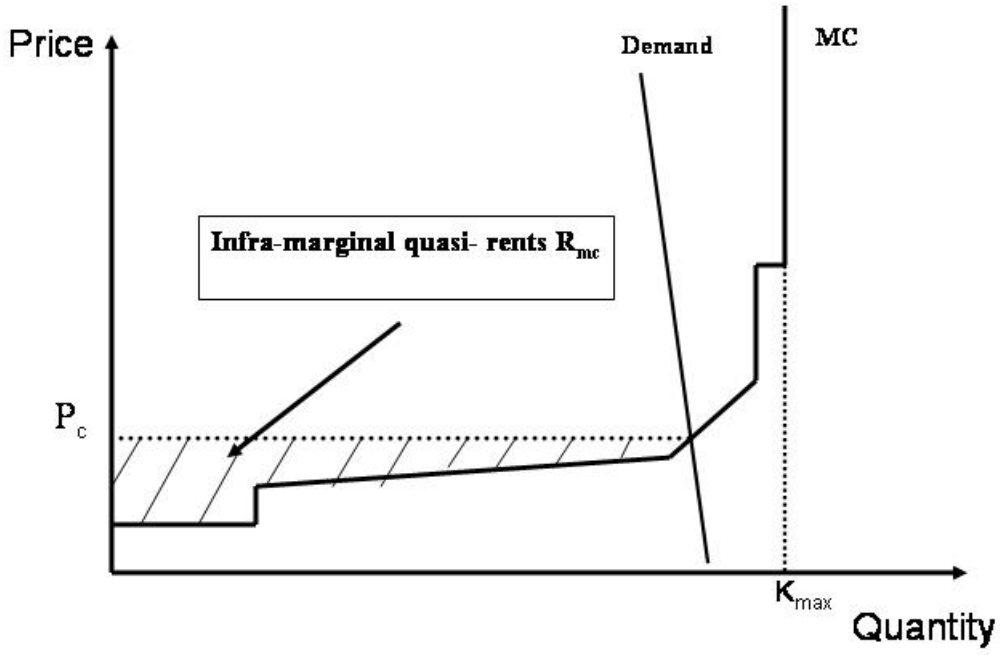


FIGURE 2

PRICING WITH BINDING CAPACITY CONSTRAINTS

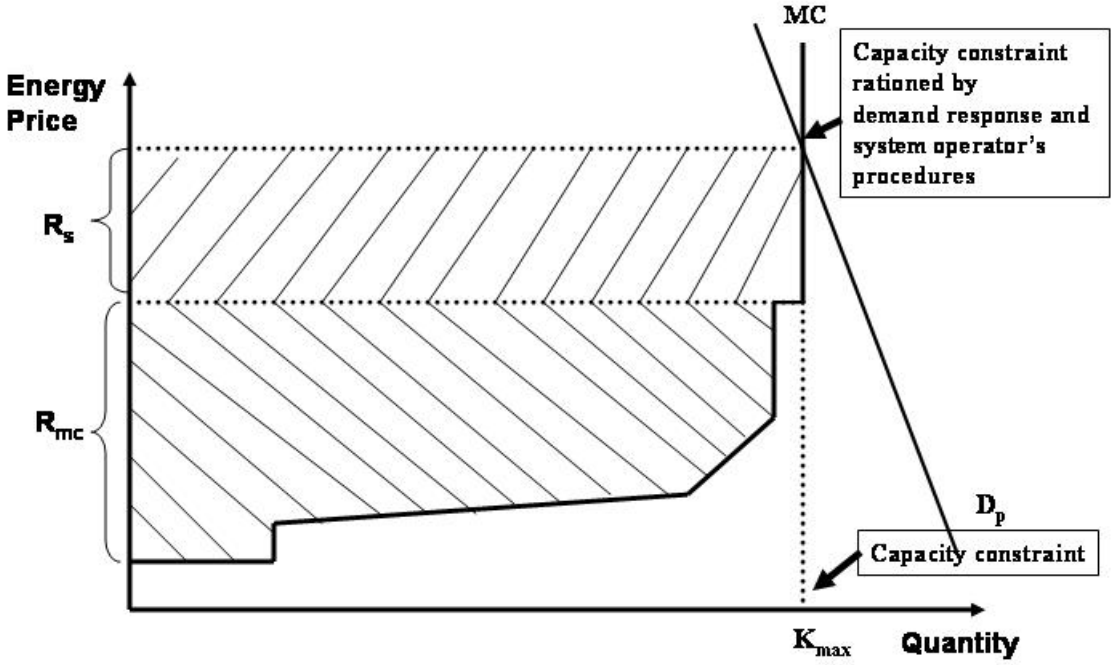
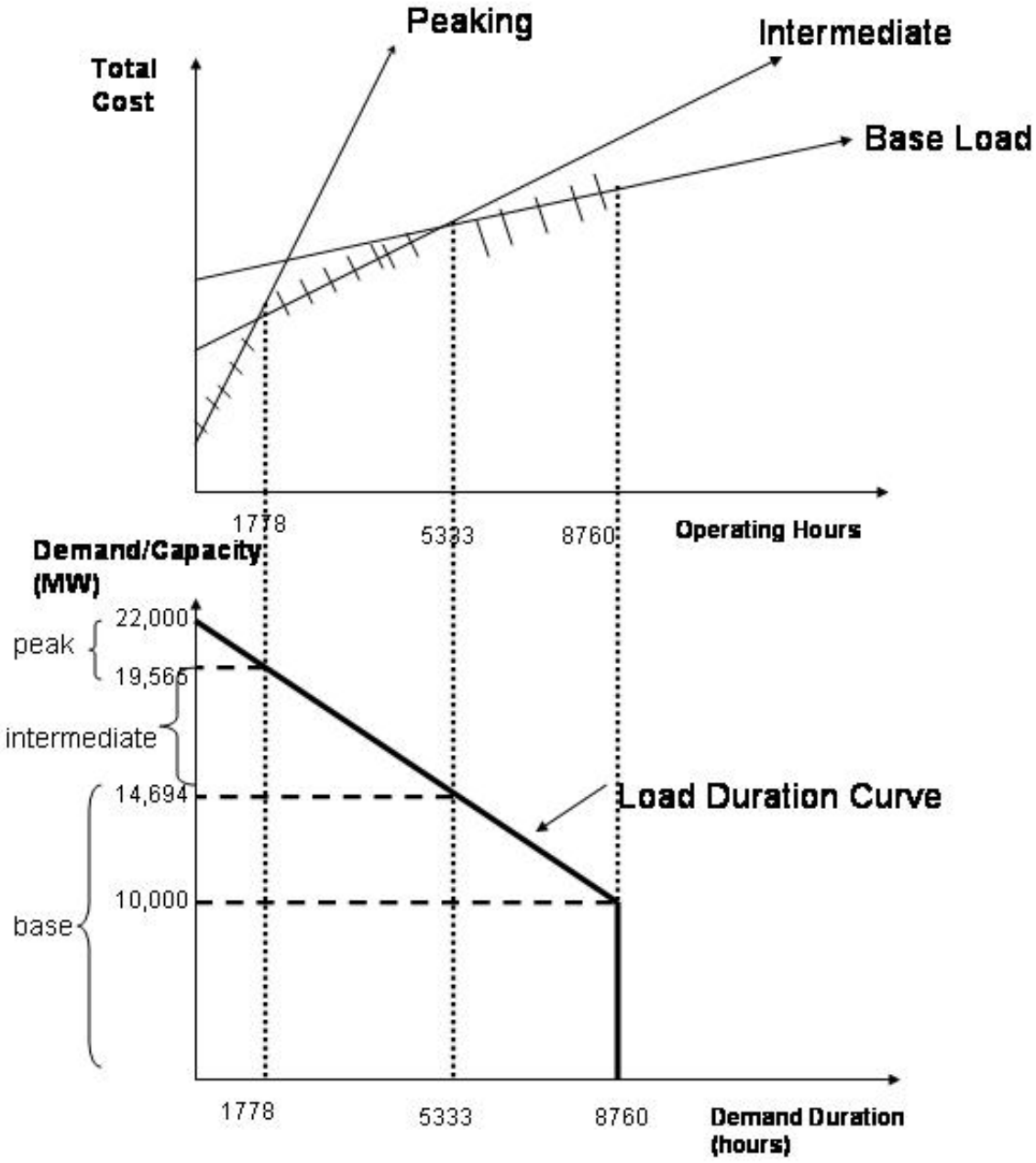


FIGURE 3


Peaking
Intermediate
Base Load
Total Cost
1778
5333
8760
Operating Hours
Demand/Capacity (MW)
22,000
peak
19,565
intermediate
14,694
Load Duration Curve
10,000
base
1778
5333
8760
Demand Duration (hours)