Ensuring Generation Adequacy in Competitive Electricity Markets[1][2]

Shmuel S. Oren[3]
University of California at Berkeley
Revised June 3, 2003

Abstract: This paper discusses alternative approaches that have been adopted around the world for guaranteeing the appropriate level of investment in electric generation capacity. We argue that long term reserves should be viewed as price insurance and be treated as a private good. However, political realities and asymmetries and distortions in risk management incentives may necessitate imposition of mandatory levels of such insurance on load serving entities. Furthermore, centralized markets may be needed as a supplement to bilateral contracting in order to facilitate efficient procurement of such insurance and to bridge the gap between the needs of generators and load serving entities with regard to duration of hedging instruments. We discuss the origins and shortcomings of capacity payments and capacity obligations and explain how long term supply contracts in the form of call options with premiums that depend on the contracts' strike prices can meet the need for ensuring supply adequacy and the financial health of the generation sector. We also outline a scheme were regulatory intervention in generation adequacy assurance takes the form of a hedging requirement imposed at the state level on load serving entities.

1. INTRODUCTION

The reliability of electricity supply has been one of the overriding concerns guiding the restructuring of the electric power industry. The slogan "keeping the lights on" has been the principal motivation for many technical and economic constraints imposed on market designs. The term supply reliability, encompasses, however, a mix of system attributes that have diverse economic and technical implications under alternative market structures. NERC (National Electric Reliability Council) defines reliability as: "the degree to which the performance of the elements of the technical system results in power being delivered to consumers within

[1] This paper was prepared under contract from the Electric Power Research Institute

[2] Some of the material contained in this paper is based on an earlier publication by the author: "Capacity Payments and Generation Adequacy in Competitive Electricity Markets", in Proceedings of SEPOPE IIV Conference, Curitiba, Brazil May 22-26, 2000.

[3] Department of IEOR, University of California at Berkeley, Berkeley, CA 94720 USA Oren@IEOR.Berkeley.Edu

accepted standards and in the amount desired". Imbedded within this definition is the notion of the "obligation to serve" which is arguably out of step with the notion of a deregulated industry with competitive supply. In fact, the concept of reliability as defined by NERC encompasses two attributes of the electricity system: *Security*, which describes the ability of the system to withstand disturbances (contingencies) and *Adequacy*, which represents the ability of the system to meet the aggregate power and energy requirement of all consumers at all times.

The notion of system security identifies short term operational aspects of the system which are characterized through contingency analysis and dynamic stability assessments. Security is provided by means of protection devices and operation standards and procedures that include security constrained dispatch and the requirement for so called *ancillary services* such as: voltage support, regulation (AGC) capacity, spinning reserves, black start capability etc.. The notion of adequacy on the other hand represents the systems ability to meet demand, on a longer time scale basis, considering the inherent fluctuation and uncertainty in demand and supply, the non-storability of power and the long lead time for capacity expansion. Generation adequacy has been traditionally measured in terms of the amounts of planning and operable reserves in the system and the corresponding loss of load probabilities (LOLP) that served as criteria for planning and investment decisions.

From a technical perspective security and adequacy are clearly closely related since a system with abundance of reserve capacity provides more flexibility in handling unforeseen disturbances. However, while a system with limited planning reserves may experience shortages it can still be operated in a secure manner while a system with ample reserve can be operated insecurely.

All the restructured electricity systems around the world recognize the need for centralized provision and control of ancillary services that are procured by the system operator either through an auction based market or through long term contracts with generators. In some cases market participants are allowed to self provide certain ancillary services but the quantities are dictated by the system operator who is also the provider of last resort for these services. With respect to long term reserves, however, there is considerable diversity in reliance on market based approaches and the debate over which is the correct way of ensuring generation adequacy is still raging. In California, for instance, where the initial market design relied on a pure market solution for provision of generation adequacy, the capacity shortages experienced in 2001 have prompted a proposal for an available capacity requirement (ACAP) to be imposed on load serving entities. Discussions concerning the appropriate form of regulatory intervention in generation adequacy assurance are also taking place in Texas were currently generation reserves are plentiful but the the California experience raises concern for the future. FERC's NOPR on Standard Market Design (SMD)[4] also recognized the need for load serving entities (LSEs) to insure the supply of power to their customers through adequate contracted provision of capacity reserves. However, the recent FERC White Paper articulating a vision for a Whole Sale

[4] FERC Notice of Proposed Rule Making, Docket No. RM01-12-000, July 31,2002.

Market Platform[5] has backed off from imposing minimum requirements for reserve margins and has recognized the States' juridstiction over resource adequacy decisions.

From an economic point of view security and adequacy are quite distinct in the sense that the former is a *public good* while the latter can potentially be treated as a *private good.* Security is a systemwide phenomenon with inherent externality and free ridership problems. For instance, it is not possible to exclude customers who refuse to pay for spinning reserves from enjoying the benefits of a secure system. Hence, like in the case of other public goods such as fire protection or military defense, security must be centrally managed and funded through some mandatory charges or self-provision rules. The resources for such central provision, however, can be procured competitively through ancillary service markets, long term contracts or other procurement mechanisms. Adequacy provision on the other hand, as will be explained later, amounts to no more than insurance against shortages, which in a competitive environment with no barriers to entry translate into temporary price hikes. Such insurance can, at least in principle be treated as a private good by allowing customers to choose the level of protection they desire. Empowering such customer choice is a fundamental goal of electricity deregulation but achieving this goal is hindered by several obstacles:

- *Technological barrier:* Enabling customer choice of generation adequacy requires deployment of metering control and communication technology that will allow differential curtailment of load when prices exceed a preselected level or will allow direct customer response to real time prices. While the rapid decline in the cost of information technologies is promissing the economic justification for direct customer load control for low end consumption levels is still questionable

- *Political barrier:* Empowering customers to make their own tradeoff between service availability and cost requires that customers be exposed to real time prices. Customers should have a default fixed price service in the same way that homeowners can obtain fixed mortgage rates but such options should be assesed a fair market risk premium. Unfortunately, electricity tarrifs are riden with politically motivated cross subsidies which distort direct assignment of costs and most public utility commissions are reluctant to embrace real time pricing.

- *Operations paradigm:* Sytems operator around the country continue to operate the system under the traditional "obligation to serve" paradigm. For example, while rotating outages are considered acceptable when reserve levels drop below a certain level (stage 3 alert) high prices (e.g. $1950 per MWh, in California during in fall of 2001, or $990 per MWh, in Texas on February 25, 2003) is not considered an acceptable reason for involuntary load curtailment. Treating generation adequacy as a private good requires a paradigm shift in system operation from an "obligation to serve" to "obligation to serve at a price".

  In an environment in which the system operations is guided by "obligation to serve at a price", the concept of loss of load probability is not well defined unless a distinction is made between probability of lost load due to

[5] FERC White Paper "Wholesale Power Market Platform", April 28, 2003

system collapse vs. lost load due to inadequate supply. It is a prerogative of consumers and producers to decide what is the appropriate level of price insurance they wish to procure and how much they are willing to pay for it as long as they are able from a technical point of view to bear the consequences of their decisions without affecting others. In the remainder of this discussion we will only focus on adequacy provision.

The traditional approach to ensuring generation adequacy in vertically integrated utilities was to build planning reserves based on load forecasts, loss of load probability (LOLP) calculation and estimates of the value of lost load (VOLL). The cost of the extra capacity was assigned to customers as a rate uplift. More elaborate schemes, attempted to allocate the cost of capacity according to time of use so that peak consumption bears a larger portion of that cost. In an ideal competitive market were prices of energy vary continuously to reflect the equilibrium between supply and demand at each moment, payment to inframarginal generators (above marginal cost) should cover their capacity cost. The question is whether market forces are indeed sufficient to provide generation adequacy given the realities of electricity markets. If market forces alone are not sufficient, as many believe, then what is the extent and form of regulatory intervention that might be needed as a transitional or a permanent measure. The primary objectives of such intervention must be to insure adequate supply of energy with minimal distortion of energy prices and investment incentives.

2. GENERATION ADEQUACY IN ENERGY ONLY MARKETS

Energy only electricity markets have been adopted in the original (defunct) California design, in Nordpool and the Australian Victoria pool. Generators in such markets bid only energy prices and, in the absence of constraints, all bids below the market-clearing price in each hour get dispatched and paid the market-clearing price. The primary income sources for recovery of capacity cost is the difference between the market clearing price and the generators' marginal costs. When ancillary services are procured separately by the system operator, as in California and ERCOT, generators can earn additional revenue by selling ancillary services, such as regulation and spinning reserve capacity, through short term ancillary service markets or long term contracts.

Economic theory tells us that in a long-term equilibrium of energy only markets, the optimal capacity stock is such that scarcity payments to the marginal generators when demand exceed supply will exactly cover the capacity cost of these generators and will provide the correct incentives for demand side mitigation of the shortage (i.e., the scarcity rend will induce sufficent demand response so that available supply can meet the remainning load). Furthermore, in equilibrium the optimal generation mix (were generators are characterized by their fixed and variable cost) will be such that the operating profit of each generator type will exactly cover their capacity costs. This optimal equilibrium mix is achieved through exit of plants that do not cover their cost and entry of plants whose cost structure will yield them operating profits that exceed their capacity costs. Figure 1 below illustrates the scarcity rent embeded in the energy price and the corresponding demand reduction (relative to the demand at marginal cost pricing) in a long term equilibrium. A shortage of capacity will will increase scarcity rents producing profits in excess of what is needed to cover the amortized capacity

cost. Such profits will attract generation expansion. On the other hand Excess generation capacity will eliminate scarcity rents driving prices to marginal cost. When this occurs, generators on the margin (like generator 5 in Figure 1) will not be able to cover their investment cost. Unless such generators receive extra revenues through some form of capacity paymens this will results in early retirement or mothballing of plants which will reduce capacity and drive prices back to their long term equilibrium level. Unfortunately, a capacity payment that will make Generator 5 content with selling its energy at marginal cost will also attarct Generator 6 to enter the market and will eliminate the incentive for demand response resulting in overexpansion at increased social and consumer cost.

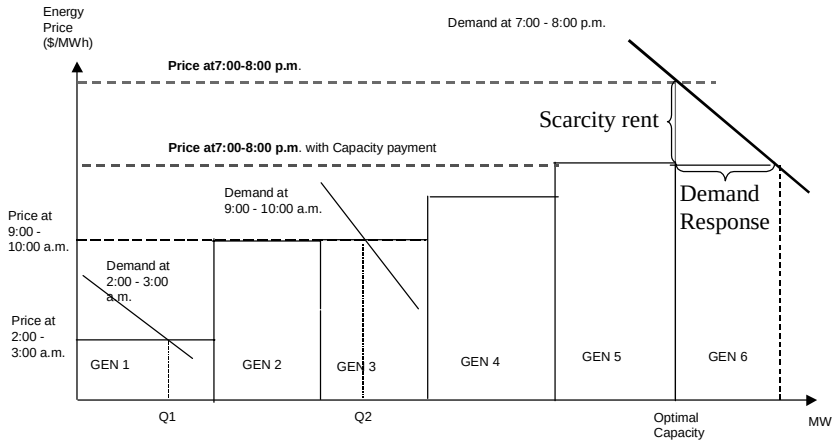

Figure 1: Optimal capacity and energy prices in long term equilibrium

The critical role of electricity in the economy and the political ramifications of widespread electricity shortages have prompted many regulators around the world to take steps above and beyond reliance on market forces in order to ensure generation adequacy. In theory, allowing the prices of energy to reflect short run supply and demand equilibrium will create market signals and provide adequate financing for proper capacity expansion. However, many regulators have been concerned that energy prices occurring in the various restructured systems are not sufficiently high to cover generators' capacity costs and to prompt adequate investment. The prevalence of regulatory intervention to suppress energy prices even when they reflect legitimate scarcity rents justifies the concern that indeed generators would not be able to cover their fixed costs through energy sales alone.

Reliance on energy prices to cover capacity costs through scarcity rents raises, many legitimate concerns. Nonstorability of electricity, demand and supply uncertainty, inelastic demand and the steepnes of the supply curve at its high end all contribute to high price volatility when reserve margins are low. While some

temporary high prices reflect legitimate economic signals that are needed in order to attract investment, they are politically unacceptable especially since it is impossible to differentiate between legitimate scarcity rents and high prices resulting from market power abuse, or from strategies such as “Hockey Stick” bidding that exploit the inelastic demand and flawed market rules. Furthermore, even if high prices do reflect legitimate scarcity rents which induce investment, sustained levels of scarcity rents while new capacity is being built will result in unacceptable transfer of wealth from consumers to producers. Such concerns have prompted the impositions of price caps and market mitigation procedures that arguably “throw out the baby with the bath water” by supresing legitimate scarcity price signals.

One could argue that the adverse effects of price volatility would be mitigated in a well functionning market by forward contracting and other risk managemnent practices. This is indeed true, however, the realities of electricity markets suggest that vertical disintegration in many restructured electricity markets and the reregulation of some segments (e.g., the retail market) has resulted in improper distribution of risk along the electricity supply chain. This misallocation of risk results in improper risk management, as was the case in California. Consequently some regulatory intervention, at least on a temporary basis, might be needed in order to achieve socially efficient risk management. Such regulatory intervention, however, requires caution since measures taken to ensure generation adequacy may have the effect of suppressing energy prices due to excess capacity or perverse incentives so that the necessity of such measures becomes self-perpetuating. This has clearly the case in Argentina, for instance, were a large capacity payment paid on the basis of generated energy induces generators to bid below marginal cost so as to increase production and capacity payment revenues.

3. THE ORIGINS OF CAPACITY PAYMENTS

Ensuring generation adequacy through capacity payments has been implemented in the UK (before the new trading arrangments (NETA)), Spain and several Latin American countries. Generators in such systems are given a per MW payment based on their availability (whether they get dispatched or not) or based on generated energy as an adder to the energy market clearing price. The capacity payments are collected from customers as a prorated uplift similarly to other uplift charges such as transmission charge. In some cases such as in Spain capacity payments are indistinguishable from stranded investment compensation, which are viewed as an additional source of revenue for the generators that is needed in addition to the competitive energy revenues in order to guarantee their profitability.

The concept of capacity payment is rooted in the theory of peak load pricing whose application in the context of electric power was pioneered by Boiteux[6]. According to this theory generation of electricity requires two factors of production, capacity and energy where the amount of energy that can be produced in any given time period is constrained by the available capacity. Consider a simple case of two consumption periods: peak and

[6] Boiteux, Marcel P. "La tarification des demandes en pointe: Application de la théorie de la vente au coût marginal", 1949, *Revue générale de l'électricité*

offpeak with two respective deterministic demand function and assume that the same fixed capacity is available in both periods. According to the basic theory, energy is priced at marginal cost in both periods and a capacity payment that would recover the fixed capacity cost is imposed on the peak-period energy users. The optimal capacity will be such that the incremental cost of a capacity unit equals the shadow price on the capacity constraint that is active during the peak. That shadow price reflects the incremental value of unserved load as measured by willingness to pay net of marginal energy cost. It is important to realize that the above approach to pricing has evolved in the context of a regulated monopoly whose primary objectives have been to recover cost and encourage consumption.

Subsequent developments of peakload pricing theory focused on two important aspects of electricity supply: uncertainty and technology mix (see Chao[7] for a general treatment of these two aspects.) The affect of uncertainty leads to redefining the basic ingredient of electricity service as energy and reliability where reliability Is manifested by LOLP calculation as a function of available capacity relative to load. The distinction between peak and offpeak than becomes a matter of degree. This perspective rationalizes levying a time varying capacity charge on all consumption and the payment to generation capacity that is not utilized for production of energy on the ground that such capacity provides added reliability. The capacity adders employed in the UK system to augment energy prices and compensate available nondispatched capacity are based on the above perspective.

Another perspective motivating capacity payments focuses on cost recovery in a system with optimal technology mix serving a load profile characterized by a load duration curve. In the following we adopt a deterministic interpretation of the load duration curve. However, a similar argument can be developed by interpreting the load duration curve as a cumulative probability distribution on load level and using average availability in determining the technology mix. Consider a set of generation technologies characterized by a fixed and variable cost per capacity increment (the variable cost defined with respect to load factor). The lower envelope of the different cost functions creates a nonlinear technology mix cost curve per capacity unit as function of operating duration. That curve can be interpreted as the system's cost of serving any horizontal load slice under the load duration curve, as illustrated on the left part of Figure 2.

[7] Chao Hung Po, "Peak load pricing and capacity planning with demand and supply uncedrtainty", Bell Journal of Economics Vol 14, (1983) pp. 179-190

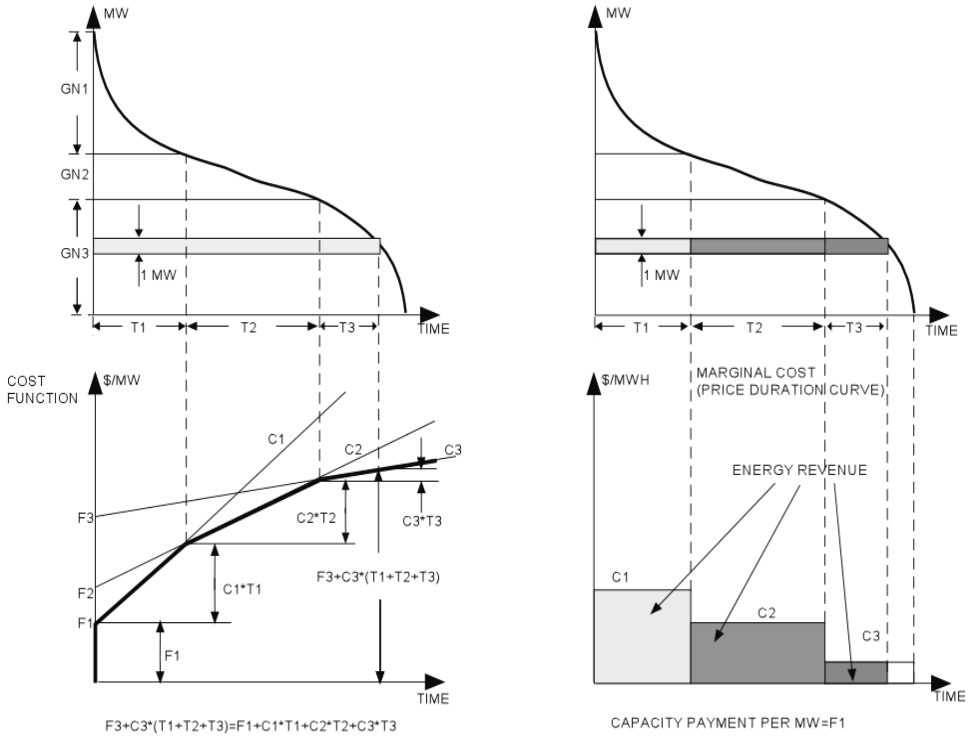

Figure 2: Recovery of geneation cost through marginal cost and capacity payments

This interpretation is the basis for Wright tariffs that price load slices nonlinearly based on load factor. In a system with coincident peaks, pricing each load slice according to the load slice nonlinear cost curve will exactly recover the total cost of generation. Furthermore, that nonlinear function coincides with the technology specific cost function in the relevant duration interval. Hence, compensating generators based on the load slice nonlinear cost curve is equivalent to paying generators their technology specific capacity and energy costs.

An alternative approach illustrated on the right hand side of Figure 2 is to price consumption and compensate generation of energy at each point in time at the corresponding marginal energy cost, that is the variable cost of the most expansive energy dispatched at that time. As we can see from Figure 2. the sum $\sum_{i=1}^{3} C_i \cdot T_i$ of marginal costs times the duration during which they are applied produces the same payment as the variable portion of the nonlinear duration-based cost function. Thus if each generator is paid the uniform system marginal cost for their energy at each point in time they end up with a shortfall in the amount of F_1, the fixed cost of the peaking technology, per each unit of capacity.

This argument rationalizes awarding generators a uniform capacity payment based on the fixed cost of the peaking technology (typically combustion turbines (CTs)) to supplement energy revenues based on marginal cost. Under optimal capacity planning the marginal cost of incremental capacity equals the marginal cost of

unserved load which can be approximated by the marginal value of unserved load (VOLL) times the probability or fraction of time that load must be curtailed due to insufficient capacity. Hence, two alternative methods for capacity payment calculation (which are, in theory, equivalent under optimal capacity configuration) are to base the payment on the cost of peaking technology (e.g. CT) construction or to use the expected value of unserved load estimated by the product VOLL x LOLP.

The need for a capacity payment to make up for generation cost recovery shortfall can be eliminated by introducing into the technology stack demand curtailment as an equivalent supply technology with zero fixed cost and marginal cost equal to VOLL. The supply curve describing cost per capacity unit as function of operating duration for the augmented technology stack starts continuously through the origin with a slope of VOLL. Hence, if we set a spot price so as to equal to the marginal cost in each duration interval, the spot price during the period where demand is curtailed should be set to VOLL as illustrated in Figure 3. Paying generators that spot price during supply scarcity periods will provide them with the same income as capacity payment. There is, however, an important difference between the two alternative forms of compensation. Capacity payments set to the value of peaking technology capacity cost fully compensates such technology even if it is idle and consequently may induce excess capacity. On the other hand paying the VOLL for energy produced during scarcity periods only compensates generators that can sell their power at that price and will hence avoid the incentive for over investment. Furthermore, capacity payments are usually paid to generators whereas curtailed load can only avoid the peak technology marginal cost of energy. On the other hand, when the capacity cost is collected by generators in the form of a scarcity spot price, the curtailed load avoids the full VOLL payment and hence such an approach incents demand side participation in shortage mitigation.

Setting the spot price at VOLL during curtailment period is a proxy to demand side bidding where true values of lost load would be manifested. Thus VOLL attempts to represent an average of the value of lost load distribution. With demand side bidding the full distribution (rather than a uniform approximation set to VOLL) is included in the supply stack. This could be depicted by replacing the straight line representing curtailment on the bottom left of Figure 3 with a concave curve whose average slope equals VOLL. The resulting spot prices during curtailment periods will at time go below the VOLL level and consequently more demand side displacement of peak generation capacity will occur.

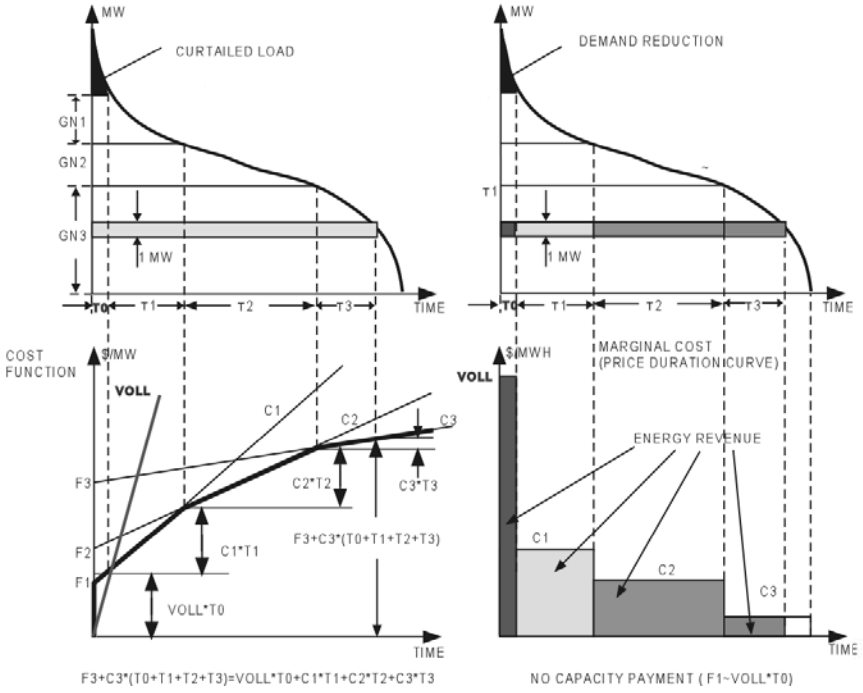

Figure 3 - Adding Demand Curtailment to the Technology Stack

In the absence of demand side bidding it is often the case that involuntary curtailments are averted by the system operator through dispatch of reserves whose energy is priced based on their marginal operating cost. Such practices give rise to an important practical question: what should be considered as a curtailment period during which the price is raised to VOLL. If the right amount of reserves were procured such deployment of reserves impacts security and that impact should be reflected in the price of energy. The use of reserves to mitigate energy shortages at prices reflecting the incremental energy costs of the reserves amounts to a subsidy between security and adequacy. This is analogous to using the army to mitigate labor shortages and charging the employers variable hourly incremental cost for the soldiers' time. A pricing scheme that would reflect the scarcity that led to deployment of reserves should augment the energy price during such periods with some prorated portion of the reserve capacity payments (of the ancillary service market) that would have otherwise been levied on all customers as an uplift. Intuitively that adder should increase gradually as more reserves procured for security purposes are being deployed to meet energy shortages and price of energy plus the adder should approach the VOLL when involuntary load curtailments are invoked.

A rigorous determination of how to set the real time spot price when reserves are being deployed would require a model that asses the effect of such deployment on system security[8].

[8] Larry Ruff, ("Capacity payment, security of supply and stranded costs"' presentation to the National Spanish Electricity System Commission, Madrid, June 7, 1999) argued that the capacity adder in the old (pre NETA) UK system was designed to accomplish this objective. In the UK there was no separate procurement of reserves by the system operator but rather all the capacity that was bid and

4. PLANNING RESERVE OBLIGATIONS AND CAPACITY MARKETS

The eastern pools in the US including PJM NYPP and New England ensure generation adequacy by imposing an installed capacity obligation on load serving entities (LSEs). Specifically, the LSEs are required to have or to contract with generators for a prescribed level of reserve capacity above their peak load within a certain time frame. The specific forms of the reserve requirement and the time frame over which such obligation are determined varies among systems and are still undergoing revisions. New England for instance, had at some point separate requirements for installed capacity specified with respect to the annual peak and separate requirements for operable capacity specified relative to the monthly peak. Formal or informal capacity markets that allow trading of capacity obligations among the load serving entities have accompanied installed capacity (ICAP) obligations. The basic motivation for the ICAP requirements is similar to the argument in favor of capacity payments. The capacity markets prompted by the obligation provide generators with the opportunity to collect extra revenue for their unutilized reserve generation capacity and provide incentives for the building of reserves beyond the reserves that meet the short term needs for ancillary services.

One could argue that if we consider generation capacity as a separate product that is needed in order to provide reliable electricity service then the supply of that product can be control either through prices in the form of capacity payments or through quantitiy control in the form of capacity obligation. The case for quantity control can be supported by the classic prices vs. quantities argument depicted in Figure 4. The basic argument is that the demand function for capacity is nearly vertical while the supply function is flat. Thus a small error in price will result in a large error in quantity so that direct quantity control is more accurate[9]. Furthermore, from an engineering and system reliability perspective the ICAP obligation insures "iron in the ground". Which is not always the case with capacity payments.

The calculation of either planning reserve requirements or capacity payments are typically based on engineering models of "loss of load probability" (LOLP) and on estimates of the "value of lost (unserved) load" (VOLL). The LOLP calculations take into consideration the quantity and mix of the available capacity in relation to the forecasted load and the probabilities of forced outages. In the original UK design capacity payments were directly computed as the product of LOLP x (VOLL-SMP[10]) and updated each half hour. In systems with mandated planning reserves, the prescribed reserves requirement are based on a threshold criterion on the expected cost of lost load given by the product of LOLP and VOLL net of energy cost.

---

not dispatched was regarded as reserves. The extent to which the load use of reserves impacts security was reflected by the LOLP calculation, which determined the capacity adder to the spot price.

[9] This argument and the illustration of Figure 4 is due to Larry Ruff.

[10] System marginal energy price

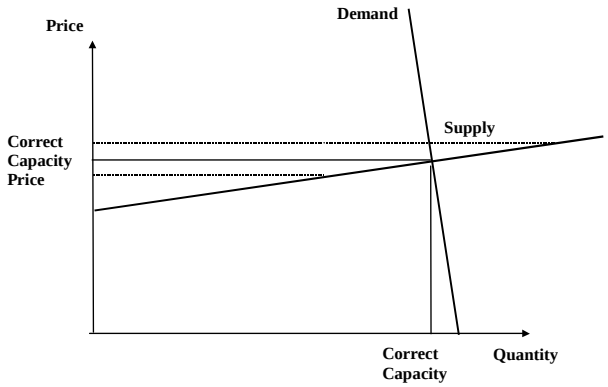

Figure 4 - Characteristic Shape of Supply and Demand Functions for Capacity

The reliance of capacity payments and capacity requirement on engineering based calculation has been criticized repeatedly on the grounds that the VOLL used in these calculations is administratively set and has no market base. The usual remedy proposed for instance by Chuang and Wu[11] is to employ VOLL figures based on demand side bidding. Further criticisms by Graves, Hanser and Earl[12] and by Hirst, Kirby and Hadly[13] point to the fact that the LOLP calculations often employ simplistic models of probabilistic failure (e.g. Poison arrivals) and do not account for more complex phenomena such as the incentives of operators to keep plants running during peak price periods. Both the arbitrariness in the VOLL and the approximate nature of the LOLP calculation are likely to result in a mismatch between energy market prices and capacity values set directly or via a capacity market induced by capacity obligations. Furthermore, as the UK experience taught us, the predictability of calculated capacity payments can lead to gaming and manipulation of the payments.

One of the fundamental problems with capacity markets is their disconnect from energy markets. The fundamental relationship between capacity and energy prices in a long run equilibrium is such that the expected social cost of unserved energy as reflected by the energy-only market prices should equal the marginal cost of incremental capacity. However, the separate capacity markets created for trading reserve capacity requirement set through engineering based methods may produce prices that are not in equilibrium with the energy market prices. For instance, overestimating the expected cost of lost load would create

[11] Chuang Angela and Felix Wu, "Capacity payments and pricing of reliability in competitive generation markets", Proceedings of the 33rd Hawaii International System Science Conference" (January 2000)

[12] Graves Frank and Philip Hanser and Robert Earl, "The death and resurrection of electric capacity pricing", Working Paper, The Brattle Group, (October 1998),

[13] Hirst Eric, Brendan Kirby and Stan Hadly, "Generation and transmission adequacy in a restructured US electricity industry", Report to the EEI (March 1999)

artificially inflated demand for capacity and result in high capacity prices which in turn will lead to overcapacity that results in suppressed energy prices and socially inefficient production and consumption. Similarly, capacity payments based on such calculations would tend to suppress energy prices to or below marginal cost resulting in excess consumption and excess generation capacity. The disconnection between capacity and energy markets is also manifested by the fact that installed capacity contracts typicall have no performence obligation requiring the capacity to be bid or to provide energy at some specified price.

Another difficulty with capacity markets relates to the time step associated with the traded obligation which is typically a month. Within this time frame both the supply and demand function for ICAP are vertical (i.e., inelastic) resulting in prices that are either very low (when supply exceeds demand) or very high (when there is a capacity shortage). Response to ICAP demand through generation planning and investment require longer term obligations.

In a market reform proposal filled with FERC the California ISO has proposed a new type of capacity obligation which tries to address some of the shortcomings of ICAP. The proposed approach is similar to the ICAP requirement in the sense that a capacity obligation is imposed on LSEs as a percentage of their forcasted monthly peak load. However, the capacity product is defined as available capacity (ACAP) which must be offered in the day ahead energy market. The ACAP product will be of various durations and can be provided either through generation capacity or physical load management. The ACAP obligations do not have however, a specified energy price ceiling which currently is by default the regional price cap mandated by FERC. That price cap, however, is subject to change [14]. Uncertainty in the price cap makes the pricing of long term ACAP contracts difficult.

5. REVISITING THE ROLE OF CAPACITY PAYMENT AND GENERATION ADEQUACY

Theoretical rationale and practical experience suggest that energy-only markets with spot prices that are allowed to reflect scarcity rents will generate sufficient income to allow capacity cost recovery by generators. Hence from a supply adequacy point of view a well functioning energy-only market can provide the correct incentives for generation adequacy. Yet there may be good reasons for some form of capacity payment and even for regulatory intervention to ensure generation adequacy. Legitimate concerns for failure of the energy markets to reflect scarcity rents or failure of the capital market to produce proper levels of investment in response to such rents may justify some intervention. In some cases regulatory intervention in adequacy assurance is needed to compensate for regulatory interference in the energy market. The supply resource stack of electricity generation in systems with significant amounts of thermal generation exhibits an inherently steep rise in cost around the capacity limit. This phenomenon combined with the typically low short-term elasticity of electricity demand tends to produce high price volatility in fully competitive energy spot markets. Spot markets that clear on an hourly or half hourly basis tend to average out some of the volatility but even in such

[14] The price cap in the west was raised by FERC from $90/MWh to $250/MWh in September 2002 and is likely to be increased eventually to $1000/MWh, which emerges as a natinal bid cap standard.

markets it may be politically infeasible to allow the energy spot prices to fully reflect scarcity rents. Consequently, energy prices are often suppressed through regulatory intervention (price caps and market mitigation) and by the market design, which in turn creates revenue deficiency for the generator that may cause insufficient investment in generation capacity. Often the threat of regulatory interference to curb scarcity rents is sufficient to inhibit capital formation and raise the capital cost for investment in generation capacity. Such interference is due to misperceptions and difficulties in distinguishing between market power abuse and legitimate scarcity rents. Thus, capacity payments or capacity obligations that stimulate capacity markets are largely viewed as remedial measures needed to offset suppression of energy prices and to ensure generation adequacy.

A useful perspective in addressing the generation adequacy problem is to view the regulatory intervention as a form of insurance against price volatility. Rather than considering the intervention as a reaction to the failure of the energy spot prices to properly reflect scarcity rents, one may regard the regulatory intervention as a proactive measure in the form of a mandatory hedge or insurance that will assure that prices stay within a socially acceptable range. Such an insurance-based view recognizes the private good nature of generation adequacy. It lays the foundation for introducing customer choice in selecting the appropriate level of price protection and for establishing a relation between the capacity payment awarded to a generator and the responsibility that such payment entails. For instance rather than setting a uniform capacity obligation or payment whose cost is evenly distributed among consumers, load serving entities, direct access customers and generators may be able to select their desired level of exposure to price risk and pay or receive an appropriate premium. Thus, generators receiving a capacity payment will guarantee the availability of their capacity to produce energy at a prespecified strike price so the capacity payment is interpreted as premium for a call option on that capacity. The higher the payment the lower the strike price and vice versa. In an ideal market where load serving entities are to free to choose the the level of price insurance they want to acquire the strike prices and corresponding capacity payments will emerge spontaneously as market based risk premia driven by the risk management preferences of the market players. However, as we explain below there good reasons for regulatory intervention at least as a transitory measure to insure that the public is protected against excessive price or shortage risk.

6. SOME CAVEATS AND IMPEDIMENTS TO MARKET BASED PROVISION OF GENERATION ADEQUACY

An important concern that is often voiced in countries where there is no well developed institutional infrastructure that can enforce financial liability of corporation is that load serving entities or generators may assume more risk than they could handle reliably. So for instance, hydro generators may oversell their water in the present market and not be able to meet their generation adequacy obligations for which they collected capacity payments through premiums on private contracts. Likewise, load-serving entities left to their own devices may not hedge their supply sufficiently in order to reduce their capacity payments and may go out of business or default on their obligation to their customers if the spot prices for electricity skyrocket due to supply shortages. We cannot ignore the reality that Us banckruptcy laws provide a de facto hedge to load

serving entities which may result in assumption of inprudent risk. This is not just a theoretical possibility, indeed the chapter eleven fling by PG&E during the California crisis and more recent bancruptcy filing by a Texas retail energy provider during the February 2003 ice storm suggest that either due to regulation flaws or by choice, load serving entities do not always manage risk in ways that are socally optimal or provide adequate protection to their customers. It is common for commercial entities involved in underwriting risk such as banks, savings and loans and insurance companies to be subject to some form of regulation that will protect the customers from default. Likewise in the case of electricity it may be necessary to set some minimum contracting or hedging level on load serving entities. The premium payment for meeting such requirements through contracting with generators will produce the capacity payments that generators need in order to insure the stable income stream for financing adequate generation investment. In exchange for a stable source of income the generators will forgo some of the opportunity to collect high scarcity rents. However, there is no need for a "one size fits all" approach that awards a uniform capacity payment to all generators and imposes a uniform capacity charge on all the loads. A market based approach, which allows parties to trade energy price risk, and investment risk through different contractual arrangements can achieve better efficiency in risk sharing and investment. Regulatory intervention can then be limited to enforcement of minimal hedging requirement and oversight of commercial liability standards and adherence to contractual arrangements.

A system of capacity payments that is linked to assumption of energy price risk can also address the problem of over or under compensation of generators based on simulated market conditions. In Colombia for instance capacity payments to generators are based on simulation results of hydro scarcity and forecasted need for dispatch of thermal plants under such scarcity conditions. Generators that are not "dispatched" by the simulation are not entitled to capacity payments although they may still be dispatched in reality whereas a generator that received the capacity payment may be unavailable.

Another problem that may arise in a market based capacity payment system concerns possible failure of the capital market to provide long term financing for generation investments at rates that commensurate with the associated risk. Such market failure may arise since supply contracts that will provide the equivalent capacity payments as option premiums are typically of short duration (no longer than five years) whereas generation investment requires fifteen to thirty years of financing. The practice of securitizing long term investment by rolling over short term contracts is prevalent in many industries (e.g. using short term savings to finance thirty year mortgages). However, lack of experience with commodity trading in the electricity industry and the perceived regulatory intervention risk (especially in developing countries) may raise the cost of capital to levels that will reduce investment below the efficient adequacy level. Capacity payments are often viewed as a means of income stabilization that would enable generators to obtain financing for adequate investment level. If this indeed were the concern that capacity payments address a more appropriate mechanism would be some form of loan guarantees by the regulator. Since regulatory intervention is one of the key risk factors

concerning investors in this business, government backed loan guarantees may inspire confidence in the regulator's commitment to uphold free market principles.

7. CAPACITY PAYMENTS AS CALL OPTION PREMIUMS

A call option is a financial instrument the right to purchase the underlying commodity at a agreed upon strike price. A system were the capacity payments represent a call option would require generators that receive capacity payments to be available to produce energy at the strike price, or purchase it and provide it at that price. On the other hand, generators that did not receive capacity payments should be allowed to collect whatever prices the market will bear[15]. The short term inelasticity of demand and steep supply curve may necessitate the setting of a price cap at an administratively chosen VOLL. That cap value will then serve as both, a penalty for unmet availability obligation and as a cap on the scarcity rents collected by generators who did not receive capacity payments. Further extension of this approach would allow generators to select among different levels of capacity payment in exchange for being available to provide energy at corresponding strike price levels, or buyout of their obligation at VOLL.

[15] Capacity payment based or call option premiums have been proposed for the Columbian electricity system in a report by Teknecon Energy Risk Advisors, LLC, February 2001.

Viewing capacity payments as premium for call options at corresponding strike prices requires the specification of contract duration. Locking in the capacity payment for a longer duration has the effect of averaging out price volatility thus, providing the security of a stable income stream for the generator and stable energy prices for the consumers. However, the argument for diversity of choices in strike prices also applies to diversity of choice in contract terms. As contracts get shorter the corresponding option premium constituting the capacity payment becomes more volatile and starts to behave as a spot market for capacity. At the limit the capacity payment becomes an energy adder, which is, indistinguishable from energy payments for dispatched generators or from ancillary services payments to generators providing spinning reserves. Ideally the capacity adder should be rolled into the energy bids and reflected in the hourly or half-hourly energy market clearing prices. When a subsequent ancillary service market exist as in California, equilibrium between the energy and ancillary service market dictates that energy bids are raised by the opportunity cost of selling capacity in the ancillary service market. Hence, the market-clearing price for reserves is a good estimate of the capacity component contained in the market clearing prices for energy. In the old UK system that equilibrium condition was enforced administratively by calculating a capacity adder based on the product LOLP x (VOLL-SMP) which is paid to dispatched generators on the top of the system marginal energy price (SMP) and to non dispatched generators that declare availability. Excess availability will depress the capacity adder but all the available capacity receives that payment regardless of the price that they bid for energy. An option premium based calculation of the capacity adder would adjust the capacity adder according to the energy price bid by the generator. Thus dispatched generators would receive an option premium based on the hourly SMP serving as strike price while generators whose bids exceeded the SMP should be paid a call option premium according to their energy bid serving as strike price.

8. A STRAW PROPOSAL FOR PROVISION OF GENERATION ADEQUACY THROUGH HEDGING OBLIGATIONS

Under this scheme LSEs are required to hold at the beginning of each month verifiable hedges in the form of forward contracts and/or call options totalling at least some predetermined percentage (say 112%) of their next month forcasted peak load[16]. Qualifying hedges must have at least two years duration with no less than one year remainning life whereas the strike prices for the call options should be at or below a maximum level set by the regulator (e.g. price cap). The requirement for long duration hedges is needed in order to attract participation by new entrants who can offer capacity beyond the existing stock and hedge their investment by selling long term forward contract or call options. One of the major shortcoming of the existing ICAP markets is the short duration of the ICAP obligations which prevent any meaningfull response by investors to high ICAP prices. Hedging obligations imposed on LSEs may be multitiered with respect to the strike prices, creating an effective demand function for planning reserves. For example an LSE may be required to hold 104% (of forcasted monthly peak load) in forward contracts and call options having strike prices not to exceed $400/MWh, another 4% with strike prices at or below $600/MWh and 4% with a strike price at or

below $1000/MWh. Hedging obligations can be met by a portfolio of contracts with generators and curtailable load resources. A limited amount of financial self-provision (backed by rigirous credit worthiness requirements) may be allowed. Under such self-provision the LSE will be obligated to absorb the difference between the market price and the strike price without being able to pass it to its customers. The determination of strike prices and the quantities of hedges at each price enables the regulator to shape the price volatility in the market. Such control is a delicate task that needs to balance the proper short-term price signal with public risk management objectives.

Call options may be procured by the LSE through bilateral contracts with generators or load, they can be self-provided by LSE controled resources (or through a financial security as described above) or they can be procured through a voluntary auction hosted by the ISO. In order to reduce the exposure to generators providing the call options and consequently reduce the call option prices, the strike price of the call option may be indexd to fuel cost. Alternatively the call options may be defined on the "spark spread" [17]. When call options are excercised the counterparty is obligated to provide the contracted power at the strike price or be liable for the difference between the market clearing price and the strike price times the called quantity. The proposed ACAP in California may be viewed as call option oblications with a strike price that equals to the price cap (which is currently rather low). The problem with the ACAP is that the strike price is not explicitly stated and therefore long term ACAP contracts are viewed as too risky on the supply side and consequently too expensive from the demand perspective.

The provision of supply adequacy through hedging obligations captures several important features. First of all this approach focuses on mitigation of price volatility as a pimary objective rather than on "steel in the ground" which is only one of the possible market responses to anticipatd supply shortages. The ability of the LSEs to meet their hedging obligations through alternative means will discipline the capacity supply and maintain equilibrium between investment in new generation, demand response and risk managment. Specifying the hedging obligation in terms of long term instruments facilitates investment sponse. Investors in new generation can raise capital by issuing the long term obligations which can be subseqauently traded among the LSEs in secondary markets. Facilitating the market for such long term hedging obligation enables reserve generation capacity to secure a stable income stream in exchange for a tangible commitment to sell energy at reasonable prices when needed. Since the LSE obligations may be adjusted mothly to reflect fluctuations in forcasted peak demand, a secondary market for call options should emerge (similar to the ICAP markets) that will enable the trading of call options among the LSEs who may wish to adjust their positions. The prices of the option will fluctuate from day to day to reflect demand for the call options. However, the generators underwriting the options are not exposed to these fluctuations. Short-term price fluctuations of long-term call options are analogous to the daily fluctuations in long-term trasury bod prices.

---

[16] The 112% figure was proposed by FERC as a minimum in the SMD NOPR (Docket No. RM01-12-000), July 31, 2002. However, the SMD NOPR recognizes that state regulator wishing to protect customers against shortages or price spikes may wish to increase the proposed percentage.

[17] Spark Spread=electricity price-heat rate adjusted fuel price.

The treasury who issued the bonds is immune to such fluctuations unless it issues new bonds or recalls existing ones.

As the market matures, individual hedging obligations may be relaxed if the market as a whole proves to be properly hedged in the aggragate.

9. CENTRALIZED PROCUREMENT

One of the major objections raised by LSEs against long term hedging requirements concerns the discrepency between the length of the required hedging contracts and the LSE's business planning horizon. LSE peak load forecasts vary from month to month due to seasonal effects. Retail competition that exists to various degrees in deregulated electricity systems adds uncertainty and variability to the LSE's peak load upon which the the hedging obligation is based. Consequently, many LSEs raised objections to having to carry two to three year hedges based on an amount that may vary from month to month. While secondary markets for hedges would allow LSEs to adjust their positions each month, the price volatility in such markets increases LSEs risk and their cost of doing business and may arguably supress retail competition. A solution that will adress the above problem without shortening the duration of the hedging contracts is to treat the hedging obligation as another ancillary service, allowing self-provision through bilateral contracts with the ISO being the provider of last resort. Under such a scheme all hedging contracts whether self-provided or centrally procured by the ISO through a periodic auction will meet the criteria outlined in the previous section. However, the cost of the centrally procured hedges will be assigned to the LSEs that meet their obligation through the ISO on a monthly basis[18].

The primary objective of the the regulated hedging obligation and centrally procured hedges is to create a backstop mechanism for ensuring generation adequacy by providing a cashflow stream to reserve generation capacity which is not capable to recover its cost due to supression of scarcity rents. The danger in instituting such a mechanism, however, is that it may interfere with the contract market and be perceived by some LSEs as an alternative to prudent risk management practicies. Long term (3 years) call options with fairly high strike prices (say 50% of the bid cap) will serve the generation adequacy objective while minimizing interference in the contract market. A high strike price will deter LSEs from leaning on the centrally procurred hedges as a substitute to bilateral forward contracting and will reduce the exposure of the ISO underwriting the centralized procurement. It is important however, to maintain sufficient headroom between the strike price of the hedging obligations and the bid cap imposed on generators. For instance if the strike price is set to the currently prevailing bid cap of $1000/MWh then that will become a defacto cap for all the generation capacity that sold call options but the cap should be raised (or eliminated) for generation capacity that is not bound by call options. The headroom between the bid cap and the strike price, effectively creates a

[18] Central procurement of long term call options (by ERCOT) on behalf of the LSEs was proposed by Reliant CO. at workshop on generation adequacy provision held by the Public Utility Commission of Texas.

two tier bid cap: a high "demage control" cap and a lower "compensated" cap. This diferentiation plays an important role in encouraging demand response and in performance enforecement, as discussed below.

It is expected that a significant portion of the hedging obligation imposed on LSEs will be self-provided through bilateral contracts that the LSE enter into as part of their regular risk management. The self-provided hedges can be backed by verifiable physical supply and resources or by verifiable demand response commitment. Alternatively the financial effect of self-provision can be replicated through bilateral financial contracts for differences (relative to the ISO prices) between the LSE's and any willing counterparty, which may or may not cover its position by selling long term hedging contracts in the ISO procurement auction.

One of the central issues in implementing the above scheme is the intertemporal allocation of the procurement cost incurred by the ISO to the LSEs relying on the central procurement for meeting their hedging obligations. It is intuitivlly clear that reserve margins that are based on an annual peak load level are more valuable (as a hedge against high prices) during high demand period when shortages are more likely. Consequently load operating during high demand periods should bear a larger share of the reserves costs. In a decentralized market for hedges as described in the previous section, LSEs will seek to reduce their holdings of call options during low load months resulting in reduced prices. Likewise, during high load months LSEs needing to meet their increased hedging obligations will cause a rise in the market prices for call options. When long term call options are centrally procured and their cost is allocated to load on a monthly basis, the itertemporal price fluctuations reflecting the supply and demand for the hedges must be computed. A reasonable allocation scheme would spread the cost of hedges over time in proportion to the loss of load probability (LOLP) which is a function of the reserve margins relative to the load level (produced by standard engineering calculations). Alternatively the allocation can be based on the probability that spot prices will exceed a given strike price.

Another issue that must be addressed under the central procurement approach concerns the penalties that should be imposed for non performance by a hedge provider. In other words, if a generator or curtailable load contracted under a call option fail to deliver energy when called how should they be penalized?. At the minimum such a resource should be liable for the difference between the market price and the strike price. However, if the strike price is close to a price cap on energy such penalty may not be severe enough since it does not reflect the difference between the value of lost load (VOLL) and the price cap. A more severe penalty that is based on estimated VOLL or forfeiture of capacity payment for several months might be more appropriate as a deterant to non-performance. Under a two tiered bid cap, as discussed above, non performing generators who are bound by a call option should be liable for the difference between the bid cap and the strike price for the undeliverd energy. In addirtion nonperformance penalties may include forfeiture of the capacity payment for the month during which the nonperformance event occured.

10. SUMMARY

The role of capacity payments in ensuring adequacy of supply can be fulfilled by risk management approaches and hedging instruments that permit diverse choices and promote demand side participation. The market should determine the value of capacity as a hedge for price risk. If capacity payment are intended to correct failures of capital markets then regulatory intervention should address directly the availability and cost of long-term financing for capacity expansion secured by short-term contracts (e.g., through loan guarantees) and focus on promoting market confidence and rules that facilitate liquid markets for energy futures and other risk management instruments

When energy markets are not sufficiently developed to provide correct market signals for generation investment, setting capacity requirements with secondary markets that enable trading of capacity reserves is the preferred approach. It is more likely to produce correct market signals for investment than administratively set capacity payments which are likely to distort energy prices and result in over investment.

A more market friendly approach that will guide markets toward prudent risk management practices is to impose hedging requirements on LSEs. Such hedging obligations can be met through bilateral trading or through centralized procurement of long term hedges. The cost of these centrally procured hedges is then allocated to the LSEs based on monthly forecasted peak load and some intertemporal allocation rule that reflects seasonal variations in the insurance value of the procured hedges. Under such a scheme incumbant generators and new entrants can secure capacity payments in the form of a premium for a long term call option they sell with a mandated strike price. The LSEs on the other hand will face a monthly prorata cost of the call options. such a scheme solves the crdit risk problem that may be faced by some LSEs if they attempted to meet their hedging obligation through bilateral contracts is likely to reduce the cost of meeting such obligations.