ELECTRICITY MARKETS AND CAPACITY OBLIGATIONS

A Report for the
Department of Trade and Industry

Prepared by NERA

13 December 2002
London

Project Team:
Graham Shuttleworth
Jonathan Falk
Eugene Meehan
Michael Rosenzweig
Hamish Fraser

n/e/r/a
National Economic Research Associates
Economic Consultants

15 Stratford Place
London W1C 1BE
Tel: (+44) 20 7659 8500
Fax: (+44) 20 7659 8501
Web: http://www.nera.com

An MMC Company

TABLE OF CONTENTS

cont/….

…/cont

DISCLAIMER

1. INTRODUCTION

This report was commissioned by the DTI as a follow-up to NERA's earlier report on security of supply in electricity and gas markets.[1] Several deregulated electricity generation markets have obliged retailers to show that they have sufficient capacity to serve their load, as a part of the market framework. This paper examines these capacity obligations from several perspectives: theoretical, practical and empirical.

1.1. Background

A capacity obligation in its general form requires that any entity that is serving load (ie a retailer, known in the US as a Load Serving Entity or LSE) must have in place sufficient "installed" or physical capacity to meet the peak load of the customers it serves plus a reserve margin.

The capacity obligation is a carryover from power pools and other regional reliability organizations. The US system is characterized by a highly integrated physical network of presently or formerly vertically integrated utilities of various sizes (with both size and geographic location being random) that cannot provide adequate reliably economically in isolation. (Reliability, as used in this document, means supply adequacy or the ability to meet total demand.) These utilities have agreed to share the consequences of supply shortages on the interconnected grid by proportionally sharing in voltage reductions, curtailments and the like without regard to the individual system that is the source of the problem at any time. As the price of entry for this reliability sharing, participants have been required to pull their own weight.

Pulling one's own weight has generally meant ensuring that each utility had an adequate amount of installed capacity in place to meet its peak load plus reserves. Over time more elaborate sharing mechanisms were sometimes implemented that adjusted capacity for quality (availability rate) and adjusted the capacity requirement for factors such as the overall load shape and coincidence with regional peak. In some markets these capacity obligations were formal and contractual and had penalties attached while in other regions the capacity obligations were informal agreements without any defined enforcement mechanism.

When competition was introduced these obligations were either dropped or formalized and capacity markets were created. As will be discussed later, the three East coast ISOs that were formal power pools developed formally defined capacity markets, while California, which had not had an effective power pool (due to the fact that it was dominated by two utilities large enough to internalise many of the reliability costs), did away with the

[1] NERA (2002), ***Security in Electricity and Gas Markets***, Final Report for the Department of Trade and Industry, Ref: 003/08 SGEM/DH, London, 21 October 2002

obligation. Experience might suggest that the East coast Pools took the more successful route, but they differ from California in too many ways to ascribe the difference in their performance solely to the retention of capacity obligations. If anyone wishes to propose capacity obligations, they must seek reasons in the fundamental economics (and politics) of electricity markets.

1.2. Basic Reasons for Capacity Obligations

Where capacity obligations exist, they are a natural outgrowth of the process which preceded a competitive generation market. Regional councils specified amounts of capacity which regulated electric companies had to build. While, in principle, regulators could overrule the decisions of these regional councils, they rarely did so in practice after the 1970′s.

To foreshadow the arguments we will be making, capacity obligation methods represent a distrust that the market can be relied upon to build the same capacity which the old system would have mandated. There is a well-known market paradigm in which competitive market pricing rewards all investment in a least-cost and diverse portfolio of generation. (This market paradigm is explained in Appendix A.) However, this paradigm relies heavily on the ability of short-term electricity market prices to soar to very high levels during a shortage, in order to remunerate investment in generation capacity that only runs at peak times, and indeed to remunerate all investments in capacity needed to meet peak demand. There are several reasons why policy makers might be concerned that this market paradigm will not function well and why, possibly, capacity obligations might be beneficial.

1.2.1. Public good arguments

The principled arguments in favour of retaining a capacity obligation have, at their core, the sometimes unspoken premise that supply adequacy is a shared good. This view is sometimes nothing more than a statement about a historical situation, but it is further justified by the fact that, in a retail competition environment, there is no way to discriminate physically between customers of a supplier able to serve load and those of a supplier unable to meet its commitments. To the extent that aggregate supply is inadequate, there will be random curtailments imposed on customers, even those who have contracted with a supplier that has adequate resources.

Implied in this statement is that demand is independent of price and that physical (non-economic) rationing is the method for balancing supply and demand in the operating time frame. To the extent that the consequences of shortages are shared in a way that has nothing to do with a customer or a supplier having arranged an adequate supply, there is a good argument for requiring a minimum level of capacity.

Electricity markets internalise some of the costs of reliability by putting a high price on peak-time injections and withdrawals of electricity that take place outside any contract. In the

UK, the charge for "imbalances" is intended to serve this purpose. However, as long as some rationing is achieved by forcible outages, rather than through voluntary load reductions in response to prices, customers cannot be sure that they derive the full value from any investment in capacity.

Under NETA, for instance, no-one yet knows how prices will behave during a capacity shortage: customers may find themselves being cut off (and willing to pay a very high price for power) at a time when their investment in generation capacity is earning a relatively low price, because they failed to anticipate the outage when submitting offer and bid prices into the Balancing Mechanism. In such conditions, customers may not be willing to invest in capacity that would be efficient according to the market paradigm.

1.2.2. Price caps (real or threatened)

Where price caps of some sort have been implemented (or even just threatened) to avoid price spikes, there is good reason to be concerned that markets might fail to remunerate (and hence to encourage) investment in sufficient capacity. An electricity system which does not allow prices to reflect true supply-demand balances will tend to underinvest in capacity, leading to unacceptable levels of reliability. If price caps are regarded as unavoidable, capacity obligations can be seen as a sensible solution.

In the US, price caps are visible constraints on the operation of many electricity markets. In the UK, no such price cap applies at present, but Offer / Ofgem has imposed price caps in the past and, under the Electricity Pool, showed an interest in investigating price spikes. Ofgem's attempt to place a "Market Abuse Licence Condition" in generator licences showed a particular concern for preventing price spikes. Although the Competition Commission rejected this condition, investors would have every reason to expect the resumption of pressure for price caps, in the event of repeated price spikes.

The economic circumstances of electricity markets make the imposition of price caps very likely regardless of a market's own history. Peak prices rise to a multiple of "normal" prices, so that any consumer exposed to wholesale market prices (which includes any consumer whose tariff can be changed at short notice) faces a choice between (1) paying a high price for electricity at these peak times; (2) investing in generation capacity or a contract to stabilise costs at peak times; or (3) lobbying political institutions for a price cap. Given that price spikes are likely to coincide with forcible outages (black outs) for some customers, high profits for some traders and generators, and accusations that generators are abusing market power, the lobbying option will appear both cheap and likely to succeed. In these conditions, investors would be right to expect that electricity markets will never be allowed to set peak prices at market-driven levels. The absence of such peak prices would constitute a serious incidence of market failure, given the reliance on peak prices to remunerate and to encourage investment in generation capacity.

Given this problem with price caps (real or threatened), it may not be possible to achieve the efficient level and choice of investment predicted by the market paradigm. Instead, policy

makers must choose between various directions, each slightly less efficient. In this context, capacity obligations may be one of a number of alternative mechanisms that improve efficiency by avoiding the problems outlined above. Below, we discuss whether capacity obligations can help in this manner.

1.2.3. Investment incentives

A second possible reason for distrusting spot markets is that spot prices may not provide a reliable indicator of the need for capacity, given the long lead time needed to construct new generating units. For instance, high spot prices may indicate a temporary lack of capacity, which is not expected to recur, or which will shortly be alleviated by new plant already under construction. In either case, it would be inefficient to start new construction projects in response to short-run price signals.

The short-run signals of the market are in fact unreliable for a great number of products for which we nevertheless allow free decisions of competitors to enter and exit. The case for imposing additional obligations on participants in the electricity market must rest on some notion that electricity is somehow "special" either because of its relatively low elasticity or because it is an essential input into a broad range of other products and services.

Here, the concern arises over the likelihood that investors will plan investment efficiently to accommodate long-run trends in demand growth, but will fail to avoid short-term problems due to unpredictable factors like severe weather, or "type faults" that temporarily prevent large swathes of generation capacity from operating. Such outcomes have been observed in several countries and cause temporary price spikes. The timing of these price spikes is unpredictable, but over the whole life of the asset they provide a key source of value that remunerates investment in generation capacity. If investors anticipate that such price spikes would (or are likely to) invoke the kind of interventions discussed above (particularly price caps), their incentive to invest in capacity will be distorted or diminished.

1.2.4. Price volatility

A third concern may be price volatility. Explicitly ordering more capacity than the market would otherwise build will tend to lower short-run price volatility as well as the average level of energy prices. If price volatility is felt to have deleterious consequences on its own, capacity obligations set above the market level will reduce the price levels for energy. Of course, the cost which society pays for this excess capacity represents an efficiency loss. Since price volatility can also be avoided by long-term contracts, it is unlikely that price volatility by itself provides any justification for capacity obligations, unless it is linked to the prospect of market failure outline above.

1.2.5. Summary

Capacity obligations emerged in the US as a continuation of previous planning techniques and are now recognised in the US as a possible corrective mechanism, required to offset the

effect of price caps on investment incentives. In the UK, no price caps apply to the electricity market at present, but there is good reason to suspect that similar interventions would be likely in the UK, both because such interventions have been observed here (and in other countries) and because the economics of political institutions suggests that such interventions will appear cheaper (to customers) than investing in more generation. In the UK, therefore, discussion of capacity obligations will focus on whether they mitigate or avoid the kinds of problems that spot market pricing would create.

1.3. Experience of Capacity Obligations

The experience with capacity obligation schemes has been mixed up to now. Where substantial excess capacity existed at the time of implementation, capacity obligations have had, as might be expected, little effect. Some plans, like New England's, were so oddly structured that they did not really amount to capacity obligations at all. The early schemes, however, have had the unfortunate tendency to replace volatility in energy markets with volatility in capacity markets. We will discuss some of the proposed measures to deal with this problem, since they would clearly be relevant to any attempt to use capacity obligations as a way of avoiding price spikes and government intervention in markets.

1.4. Outline of Report

We begin, in section 2, by setting out in more detail the theoretical arguments for and against capacity obligations. Section 3 will enumerate the most important practical considerations in implementing a capacity obligation.

Section 4 will then discuss the experience with capacity obligations in the US, particularly focused on the three large northeastern markets: Pennsylvania-Jersey-Maryland (PJM), New York and New England. Section 5 speculates on the future of US capacity obligations particularly in the proposed Federal Energy Regulatory Commission's (FERC) Standard Market Design (SMD) initiative.

Section 6 sets out our conclusions, in terms of the design requirements for any capacity scheme and the elements likely to ensure successful encouragement of adequate investment in generation capacity.

2. CAPACITY OBLIGATIONS IN THEORY

2.1. The Role of Contracts

When customers are worried about possible variations in the cost of supplying their needs, the normal reaction is to contract for future demands with some supplier or set of suppliers. Armed with ensured demand for his output, the supplier can dedicate facilities (including incremental ones) to this customer.

While such contracts can serve to hedge the financial risks of delivery in electric markets, they cannot be used to hedge the risks of actual delivery. Unlike other markets, the output of a particular generator cannot be sent to a particular consumer: the reliability of the electricity system is shared. If other industry participants have not built enough generation capacity, buyers with a long-term contract may still suffer from power cuts imposed by the system operator.

Thus, reliability of the electric system is an attribute of the entire system.[2] The reliability of the system depends only on the aggregate supply-demand balance. Like any jointly supplied product, system reliability is therefore subject to free rider problems. Mechanisms must ensure that individual consumers cannot reap the benefits of the reliability-purchasing decisions of others – otherwise, too little reliability will be supplied.

Many markets have long-term contracts between customers and producers which help ensure economic returns to the producers, particular where the capacity to supply cutstomers is long-lived. The most prominent case is where the output of the plant is tailored to the specific customer, ***e.g.*** a mine-mouth coal plant. Transportation of the coal would be too expensive to allow any other customers for the coal, so a long-term contract is required to induce dedication of the facilities.

Most markets have no such guarantees, because dedication of the facilities to a particular customer is fairly rare. There is no long-term contract between a consumer and a gasoline refiner, even though the investments in gasoline refineries are similar in capital intensity and lifetime to an electricity generating unit. Thus, we must carefully explore what makes the electricity system different in order to justify these obligations.

---

[2] By the "entire system" we mean a given market area with a uniform price. Currently, this definition would apply separately to England and Wales, as opposed to Scotland. The size of these market areas will change from time as time as constraints on the transmission system serve to separate prices between areas. The introduction of BETTA, a British electricity market, might bring Scotland into the same market as England and Wales, but the definition ought really to depend upon the incidence of transmission constraints. Capacity in Scotland might not be able to meet load in England and Wales at all times, suggesting that obligations should still be defined separately.

2.2. An Energy Market Paradigm

First, consider systems with no capacity obligations at all. Consumers of electricity simply pay the spot price for energy (meaning the price per kWh of electricity actually taken). In such a system, we might solve the problem of free-riding by requiring all consumers to pay what the market will bear to avoid being forcibly cut off. The free-rider problem is then avoided by having people pay for unreliable service at a common price. If a particular consumer has a contract with a generator, the risk of paying that common price is shifted back as an opportunity cost to the generator. The decision to construct new capacity is left to each potential generator's judgment as to the state of the market at the time his capacity will come on line.

If bidding to supply power is competitive, where do the signals to construct new capacity come from? When the available supply falls short of the demand level, the theorem that price equates to the marginal cost of the last supplier no longer applies. At that point, it is competition between buyers which determines price. In that case, the willingness to pay of the lowest-valued electricity uses determines price.[3] For the most expensive (marginal) units on an energy-only system, these moments of shortage are the only times in which they can pay for their costs over and above their marginal operating costs.

When buyers compete not to be denied power, the spikes in price can be quite large. The average value of lost load has been estimated to be in the range of $5.00-$6.00/kWh for US consumers (equivalent to £3-4/kWh, compared with normal off-peak electricity prices in the range 10p-30p/kWh). Given average annual capital costs of a peaking turbine in the range of $30/kW-year, a peaking turbine would require around such price spikes to occur for 5 to 6 hours per year ***on average*** in order to recover its total costs, including an economic return on invested capital.

Critical to this solution to the capacity expansion problem, therefore, is that prices are allowed to float freely. While these price spikes do not really raise the long-term expected (ie, average) cost of power,[4] they can cause intense political pressure. This is particularly true if, due to an unanticipated increase in demand, the high prices persist for appreciable periods of time. Although the market described above would need 5 to 6 hours of high prices on average, such hours are likely to be concentrated in specific years when capacity is short, separated by several years when no such price spikes occur. If shortages occur once every 5 years, for instance, the shortage year would need to experience peak prices for 25-30 hours in total, ie about an hour a day for one month.

---

[3] The same is actually true in the more typical circumstance of adequate capacity. The difference is that the marginal willingness to pay and the marginal cost of additional energy are the same. When supply falls short, there is a gap.

[4] If the expected price of power, including price spikes, exceeded the cost of entry, there would be an unexploited profit opportunity.

Adequately implemented demand response can reduce the level to which prices rise when there is a capacity shortage, albeit the prices must rise to these levels for more hours to remunerate capacity. However, any prices that are demonstrably well above the short-run marginal cost of generators lead to a concentration of capital recovery in a short period, which creates the illusion of excessive generator profits. Until the day that such profits are acceptable, price caps are often regarded as a political necessity.

2.3. Effect of Price Caps

In response to this pressure, and in the absence of substantial amounts of price-responsive demand, almost all jurisdictions have mandated some sort of price cap. Whenever prices are capped (except where those caps precisely correct for true market power, which is unlikely) the imposition of lower average prices reduces the incentive to build new generation units. The lower the price caps, the more likely that incremental construction is discouraged and the more likely that reliability will suffer. In the new equilibrium, reliability is lower and the cap is reached more often. The system balances by imposing more forced outages and by building less generation capacity. Such an outcome is unlikely to satisfy consumers in the long-run.

A system with uneconomic price caps (***i.e.*** those which do not reflect the prices which a competitive market would set) will not build enough capacity without some external compensating force. The best case for capacity obligations is that they present such a compensating force in the presence (or threat) of such caps. By obliging customers to pay for more capacity than producers want to build on their own to serve the market, the additional source of revenue which the artificial price caps has removed from the market can be restored.

In addition, leaving reliability adequacy for the market to sort out leaves the regulator with no guarantee that adequate capacity will be built. Cycles of "boom and bust" are present in many capital-intensive industries. Getting the level of capacity right "on average" may well be politically unpalatable. In the old regulatory scheme, the regulator had someone who was demonstrably to blame when forced load shedding occurred. Under an energy-only scheme, no one can be held responsible. Thus, a capacity obligation scheme is a response to the inherent lack of responsibility for aggregate indicators in a market (although it seems to treats "booms" and "busts" asymmetrically). When there is a shortage of capacity, it is possible to pin the blame on the particular retailers that are failing to fulfil their obligation. There is no direct economic value in this ability to blame someone for past failures, but there may be important political consequences and, just conceivably, an incentive for companies to avoid blame in order to preserve their corporate reputation.

In principle, electricity markets should rely less (or not at all) on the use of forced load shedding, because short-term prices provide an alternative method of rationing. As the price rises, customers will have an incentive to reduce their own load voluntarily, if they are exposed to the short-term market price, or if they have interruptible contracts with retailers

who are exposed to those prices. Ultimately, the expansion of demand involvement might lead to a situation where a shortage of capacity is resolved entirely by customers reducing their demand voluntarily, instead of system operators cutting off customers. In such cases, a shortage of capacity may become evident from the high level of prices, rather than forced outages.

However, some forced outages may still be required (1) if capacity shortages are localised in areas where there are few price-responsive customers and (2) if the shortages occur to quickly and unpredictably for customers to react. A period of low prices, with low volatility and relatively little difference between peak and off-peak prices (as observed currently under NETA) provides just as little incentive for customers to fit load management technology as it does for generators to build new capacity. As a result, demand management may not develop spontaneously in response to short-term price signals. Indeed, if customers are exposed to the short-term price (as required for demand response), they may complain about the high prices or the frequency of contractual interruptions (as opposed to forced curtailments). Consequently, demand management may not develop rapidly enough to avoid price spikes and, when they occur, pressure for price caps will still emerge.

2.4. Reducing Price Volatility by Encouraging More Investment

Electricity spot prices are inherently volatile. While volatility ***per se*** is of little concern to most consumers, for whom the highs and lows of prices will tend to even out, the political pressure from consumers attendant on the periods of high prices are not met with corresponding credit in periods of low prices. Thus, there may be important political reasons to attempt to lower price volatility.

If excess capacity is constructed, spot energy prices will fall. Capacity obligations might then be viewed as a method of maintaining a level of capacity ***higher*** than an efficient free market would have sustained, thus lowering spot energy prices and relieving the transient pressure from price spikes. Of course, such methods do so only by raising the price paid for the capacity itself: faced with lower prices for energy, generators will demand higher prices for the capacity in order to recover their costs.

The standard method of lowering volatility is by contracting between buyers and sellers. The buyer gets some certainty over the price he will pay and the seller gets some certainty in the price he is paid. Both parties can concentrate on the level of price and ignore volatility in the spot market. However, volatile energy prices may lead potential producers to charge very high prices in order to cover their risks in long-term contracts. Capacity serves as only a partial hedge against high energy prices: equipment can break down, forcing the generator to purchase energy when prices are high. Second, foregoing the high profits to be earned in times of shortage can form a large opportunity cost to contracting which the generator must recover.

Some capacity obligations seek to deter the high price episodes required under an energy-only framework by attempting to ensure that prolonged capacity shortages never occur in the first place. Loads are obligated to specify ("nominate") in advance the particular units which will serve their load with an appropriately determined margin of reserve. Thus, capacity obligation plans avoid the free-rider problem associated with the joint provision of reliability by forcing all loads to pay for their share of reserves up front. In the US, regulators usually mandate such system to provide just the capacity which the old system of regional councils provided.

2.5. Designing a Capacity Obligation

A capacity obligation scheme requires a number of interventions by government or the regulator (or a body designated by government to implement the capacity obligation, like an Independent System Operator) to function.

- First, estimates of the ***future loads of all parties*** must be obtained.
- Second, the ***timing*** of the capital adequacy test must be set, meaning both the frequency with which capacity adequacy is reviewed and the future period to which it applies.
- Third, standards for ***required reserve margins*** must be established.
- Fourth, standards for the ***accounting of capacity*** must be promulgated.
- Fifth, ***operational rules*** must be created to ensure that the capacity promised by generators is actually forthcoming when needed.
- Sixth, ***penalties*** must be established for failure to meet the obligations.

In theory, if all six of these determinations were made optimally, a capacity obligation plan would achieve an optimal level of investment in new plant. The demonstration of this theorem is beyond the scope of this paper, but it is important to see that, just as a centrally planned economy could - in theory, given enough information - yield an optimal resource allocation, so could a capacity obligation scheme yield an optimal construction pattern. The analogy is telling in practice, however. Each of these standard-setting decisions, however, takes a decision otherwise left to the individual participants in the market and forces a government body to substitute its judgment for that of the market. Since it is precisely the fallibility of centralised decision-making in general which has led to the decision to create a competitive generation sector in the first place, we need to explore what possible improvements are lost through such a scheme. Thus, the gains, if any, of such a plan need to be carefully assessed.

Capacity obligation schemes should therefore always be seen as a "second-best" alternative, required (if at all) because the "first-best" option is unavailable. In the presence of price caps, there is insufficient incentive to construct capacity to get reliability to satisfactory

levels. No one has any incentive to do this on his own, no matter how great his desire for reliability, because he has no way of capturing the value of the increased reliability.[5] Capacity obligation plans deliberately spread the burden of increasing reliability beyond the level which the price-cap restricted energy market would support.

Thus, the gain from a capacity obligation plan would be increased reliability of the electricity system. The costs of the plan depend on the specific implementation of the six functions above. These costs come in two forms.

- First, poorly implemented plans may not in fact lead to the construction of any additional capacity at all. If, for example, penalties for non-compliance (or a cap on capacity prices, which amounts to the same thing) are set too low, those bearing the obligation will simply pay the penalties (or capped prices) and continue to endure sub-optimal levels of reliability.
- Second, plans may demand high levels of reliability which emphasise the damage (to government agencies held responsible) due to capacity being insufficient and ignore or downplay the costs (to consumers) of building extra plant. Again, such consequences may be an inevitable outcome of a second-best market, but the costs to consumers are not trivial.

In the next section we explore the practical implications of these six determinations which the regulator must make to implement a capacity obligation plan. In examining these considerations, we consider both types of costs: those which undermine the goal of the obligation altogether, and those which achieve it at an uneconomic cost.

[5] Unless a consumer dedicates local backup generation to serve his or her own needs. Beyond certain applications, such as hospitals, which have extraordinary needs for reliability, such redundancy is highly uneconomic.

3. THE PRACTICAL DESIGN OF CAPACITY OBLIGATIONS

3.1. Loads

Making each load responsible for providing sufficient capacity to supply his demand requires, to begin, a division of all the loads into aggregates over which the capacity obligation will operate. It is clearly impractical for every 2 kW household to sign contracts nominating 2.4 kW of load to serve it. Market areas are divided between load-serving entities (LSEs) who bear the responsibility to find adequate supplies. Where these LSEs largely consist of the old monopoly distribution companies, this division is fairly easy to do, but if retail competition makes substantial inroads, it may be difficult to figure out just whom a particular retailer is serving. In the UK, the process of retail competition already requires careful accounting for supplier-customer relationships, for billing purposes, so this hurdle has already been crossed. However, the system is not problem-free, and it was not designed for use in identifying capacity obligations.

The capacity obligation is an ***ex ante*** obligation – the LSE must nominate its capacity in advance of the actual load being delivered. What happens to customers who shift suppliers after nomination is an important part of the rules. If, after nomination, a customer moves from one LSE to another, the new LSE will be short of capacity. If the rules are not properly structured, LSEs might be dissuaded from paying high prices for capacity, because including that cost in retail electricity prices might enable LSEs who had no capacity costs at all to capture these customers (at least for a year) leaving the original LSE without enough customers to pay for its contracted capacity. There are several solutions to this problem which depend on the specific implementation of the obligation, as discussed below.

Linking loads to suppliers is only the first step. Next, those loads must be quantified. The estimation of future loads is a task which regulated utilities have traditionally performed, with varying degrees of success. Three new problems arise, however, with the estimation of future loads in a capacity obligation framework.

First, if the former utility is not performing the load estimation function in the same way they did before (either because the loads have moved to another entity or because the estimation process is carried out by another body, e.g. the ISO) then the "tried and true" methods of load estimation may not be available. New methods will have to be audited somehow by the regulatory authority.

In this respect, the UK has already made the necessary adjustment, since NGC has procedures for forecasting total demand within the competitive market. However, NGC's long-term forecasting procedures have never before been used to drive obligations on others. A UK mechanism might also use the standard load profiles that have been adopted for non-half-hourly metered customers, for instance in order to convert data on total electricity usage by customer type into a forecast peak demand by customer type.

Second, the incentives of the estimating body need to be carefully considered. Regulated utilities had no incentive to understate demand.[6] If the LSE is allowed to estimate its own demand, however, it has every incentive to understate that load unless penalties are put in place for understatement. If, by contrast, an outside party estimates the loads, it will have little incentive to do a good job, unless its forecasts are compared with reality. This is a difficult problem to overcome, since actual peak demand can differ from forecast levels for many unpredictable reasons. This problem applies to peak demand forecasts made by NGC.

Third, errors in estimated load undermine the benefits of the system. Forecasts of aggregate changes in annual peak load have historically not been very accurate. Further, traditional systems of monopoly regulation offered no real benefit for estimating load more precisely rather than less precisely. This is not surprising, since these estimates depend on many factors which there is no particularly good way to reliably forecast today: macroeconomic trends, trends in energy intensity, and migration patterns. Part of the purpose of a capacity obligation scheme is to deal with the recognition (explained above) that price spikes will occur unpredictably.

Reliance on false forecasts that underestimated future demand would also undermine the entire purpose of the capacity obligation program, leading to exactly the reliability losses and price spikes that the scheme was meant to avoid. In the current context, the use of capacity obligations to increase investment may be intended to restore the "efficient" level of investment in capacity (and hence the efficient level of forced outages). However, government agencies (including some in the UK) may see such obligations as a way to promote ***additional*** investment in capacity (and hence to ***reduce*** the level of forced outages below the efficient level).

3.2. Timing

Capacity obligation methods require LSEs to nominate sufficient capacity. A critical question is how often they must make such declarations and how far in advance these commitments must be. The closer to real time that capacity is nominated, the surer one can be that capacity will be adequate on the day. Demand can be more precisely estimated closer to the actual time. Short-term forecasts can also take into account the up-to-date status of particular generating units, whose capacity may change from time to time.

On the other hand, the problem with very short-term assessments of capacity adequacy is that, if there are shortfalls, very little can be done. New capacity takes a minimum of two years to construct, and can take up to six years in some locations. Unforeseen events which lower the supply-demand balance below acceptable levels cannot be corrected. What then

[6] Indeed, if one believes in Averch-Johnson effects, they had incentives to overstate demand in order to allow new plant construction at a faster pace.

happens depends on the penalties for non-compliance, to be discussed below. At the simplest level, if those who cannot meet their capacity obligations were denied the right to consume beyond their demonstrated capacity, then the system would in fact always be at the margin specified. But no system proposes this because the goal is not to apportion shortages fairly, but to share out the costs of achieving a certain reserve margin. Those who fail to provide adequate capacity must pay a penalty rate, instead of doing without power.

This penalty rate becomes a de facto cap on the price that anyone will pay for capacity in the market. It also becomes the source of price volatility associated with capacity obligation schemes. If capacity is abundant, competition between those anxious to supply capacity to the market will reduce the equilibrium to something close to the fixed operating and maintenance costs of a peaking turbine, with no contribution at all to the capital cost of the generator. When capacity is tight, there are no marginal sources, and price tends to rise to the penalty price. Oscillation between these two regimes and the two associated prices then becomes the only signal sent by capacity obligation schemes.

This observation confirms that there is no way to avoid incurring the full costs of capacity if consumers want adequate supplies, and also that the value of capacity is a very volatile factor. The only way to avoid the problem of volatility (and the associated pressure for price caps that depress total returns) is to spread the remuneration of capacity over more periods by deliberately moving away from short-run marginal pricing.

In practice, the volatility in observed spot prices for capacity may substantially overstate volatility in the actual market price for capacity. Long-term contracts between LSEs and suppliers are not reported. To the extent that such bilateral contracts dominate the market, wholesale prices paid on average may be much smoother than the observed spot prices. However, in a system with retail competition and short-term contracts (ie, revisable tariffs), even short-term increases in wholesale market prices can rapidly feed through into retail electricity prices.

There is an alternative view for which support is growing. It has been proposed but not yet been implemented anywhere. This view says that the period over which a capacity obligation applies should be aligned with the time it takes to construct new units. In theory, this approach means that LSEs are not helplessly condemned to paying the penalty price until new capacity is constructed. By allowing the LSE to contract for the construction of new capacity which will come on line in the timeframe required, it is hoped that the LSE will be able to respond to the signal. However, the penalty acts as a cap on the price that the LSE will be willing to pay for capacity and, if set too low, will discourage investment.

The incentive to respond to such advance signals (presumably) requires that the LSE is able to avoid the penalty by contracting for capacity. Such a system would entail a two-stage test: the first test would establish a demand forecast (say) two years in advance, to establish whether the LSE was short or not; and a continuous daily test in the period up to real time would then check whether the LSE had secured sufficient capacity to supply the original

forecast of demand (even if subsequent forecasts, or actual demand, were lower). The LSE would only suffer a large penalty for being short of capacity if it failed to avoid the daily penalties.

The main problem with setting capacity obligations in advance is the difficulty of adjusting them when consumers switch from one retailer to another. Several schemes in the US have dealt with this problem by assuming that customers carry the obligation with them, so that retailers have to trade capacity in daily markets as they win and lose customers. This scheme allows the continuous monitoring of capacity to keep in step with the retail market, but it exposes retailers to the risk of short-term price volatility in daily capacity markets, to the extent of their daily net gain or loss of customers.

3.3. Reserve Margin

In the United States, reserve margin calculations have been the responsibility of the regional councils of the North American Electric Reliability Council (NERC). The putative standard was that generation should be adequate to provide no more than one-day-in-ten-years of lost load due to generation insufficiency. In truth, the methods used by NERC generated reliability considerably in excess of that level, but we presume that something like those levels are part of any capacity obligation program. Thus, a successful capacity obligation scheme would preserve the traditional reserve margins. This, indeed, is their major ***raison d'être.***

The problem here is that one of the potential benefits of a competitive generation sector is being sacrificed. Ever since the work of Telson (1975), there is suspicion that the capacity reliability of the electricity system has been too high.[7] At the margin, the losses from unreliability of the system to those whose power is cut off should just balance the cost of the peaking capacity which would alleviate that shortage. Substantial evidence exists that US consumers have paid more for their reliability than the benefit they are in fact getting from it, with a result that electricity consumers in the US spend about $2.5 billion per year in excessive reliability. One of the benefits of a deregulated system (without capacity obligations or payments) is that the balance between the marginal value of reserves and the marginal cost of reserves would be left to the market. Capacity obligation methods, in specifying the reliability which the system must meet, will lock those inefficiencies into place.

The estimated annual cost of excessive reliability in the US translates into about $9 per head of the population, or about £6 per head. For a country the size of the UK, this figure would translate into total annual costs of £370 million. However, the cost per head would most likely be somewhat smaller than in the US, because of the lower level of electricity

[7] Twenty four hours in every ***ten*** years allows about half the outages identified as optimal above, where we suggested 25-30 hours in every ***five*** years. However, the difference may be due to different estimates of the cost of a peaking generator.

consumption per person. At 5.5 MWh, consumption per head in the UK is about half the US level of 12.6 MWh,[8] which suggests that the equivalent annual cost of additional reliability would be only £3 per head for the UK.

3.4. Defining What Counts as Capacity

Not all capacity is created equal. The definition of capacity offered by a generator and purchase by a LSE in fulfilment of its obligation requires a strict procedure to cope with a number of distinct problems.

1. The maximum output of a unit is an economic determination. Plants can always be run a little harder to yield a little more output, albeit at the risk of catastrophic failure.
2. Many units, particularly hydro, have a limited amount of energy to offer within a certain period. There is no guarantee that the maximum energy available from the unit will be available at the time the system needs it.
3. Renewable technologies like solar and wind can only promise power at the peak (or indeed whenever needed) on an uncertain basis: Only if the sun is shining or the wind is blowing will they have power to offer the market.
4. Transmission limits may hamper the ability of units to meet load. If 2000 MW of generation is on one side of a 1000 MW-limited interface, then only 1000 MW of that capacity can be counted towards the capacity of obligation of those on the other side of the interface. Some method must be used to allocate that capacity among those participants.
5. Certain generators may be able to provide to more than one market area. Capacity obligations in both regions must be audited to ensure no double counting.
6. The units must actually be capable of running. Thus, annual testing regimes need to be implemented to ensure that the named capacity is actually available.

In the US, such problems must have been solved by existing capacity obligations. In the UK, similar rules are already required for all electricity contracts that include a payment for available capacity and can be adapted from such sources. The resulting scheme is, however, bound to be complex, as it is never simple to define capacity.

3.5. Calling on the Units Nominated

The capacity obligation is imposed on LSEs, who fulfil it by building or contracting with generators. These obligations must then become an obligation on the generator to generate power when the system needs it. Otherwise, the obligation is meaningless. Thus, if the

[8] Figures relate to the year 2000. See IEA (2002) ***Electricity Information***, pp 679, 680, 699 and 700

generator is exporting power to another system at that time, the generator must be prepared to cancel the contract to export, bearing whatever penalties are necessary to redirect his capacity to the market. The rules must set out what penalty a generator faces for refusing to behave in this way, and how the generator will be paid at the moments his capacity is allocated to fulfilling his obligation.

In the UK, the equivalent would be an obligation on generators to make all capacity available to the Balancing Mechanism if it is not being used by contracted customers.

This obligation does not offer any protection against short-run price volatility, since the obligation is not linked to the offer of energy at any particular price. Subject to energy market price caps, generators can offer their plant at any offer prices they like. As a result, when there is a shortage, energy market prices will rise as high as the cap (or customer willingness-to-pay) will allow.

Unless the economic incentives are properly aligned, there is no particular assurance that capacity will be maintained so as to ensure that the surplus capacity is available when needed. Many plans have coordinated maintenance scheduling. Depending upon the degree of central control over maintenance scheduling, this aspect again cedes a potential advantage of an energy-only system. Innovations in maintenance scheduling which might have improved the operation of electricity markets will be lost to a more centrally planned scheme of regulating maintenance periods.

3.6. Penalties

Setting up a set of rules requires a set of compliance penalties. If penalties are set too low, the obvious problem is that LSE′s will simply ignore their obligations and pay the penalties instead. This raises money for the ISO but does nothing to address the problem of capacity adequacy. Recognizing this, the tendency is to set a high penalty price, certainly one in excess of the levelized capital cost of peak generation capacity. Thus, an LSE should always prefer to pay for a peak generator rather than pay the penalty for being short of its obligation. (If the LSE expects to take a substantial amount of energy from the capacity, then other forms of generation would be cheaper still. See Appendix A.)

In the short run, however, there may be no physical way of making the necessary capacity available. In this case, the LSE will be forced to bear a penalty price set intentionally high. If energy market prices are also high at this time, the revenue gained from the penalty will not serve any worthwhile purpose. The value of a capacity obligation depends on the presumption that energy market prices are subject to (or threatened with) price caps.

While capacity obligations can be useful to ensure adequate capacity in the presence of price caps, they can only work if the penalty prices are not only set high enough, but stringently applied. Again, in periods of capacity shortage for any reason, the average price of power

will rise to reflect the penalties which LSEs are paying. If these payments are also capped, the problem of inadequate capacity will reoccur.

The proper level for a penalty is one which approximates the social cost of the outages which the additional capacity would have forestalled. If the LSE chooses to pay the penalty, then it obviously cannot find capacity at less than the social cost, so its decision to pay the penalty is optimal. Calibrating the proper level is still a matter of art, requiring estimates both of the value of lost load and the loss-of-load probability. The loss-of-load probability depends upon the timing of the calculation. In real-time, capacity is either short or not, loss-of-load-probability is ***either*** 1 ***or*** 0. The required penalty is only invoked in the former case, so it has to be similar to the value of lost load. In advance, the loss-of-load probability is always ***between*** 1 and 0, so the associate penalty can in principle be somewhat lower. However, the appropriate value depends upon the nature of the scheme.

4. ASSESSMENT OF ACTUAL EXPERIENCE

In the United States, restructured and integrated electricity markets have so far only existed for a few years, in four regions: California, PJM, New England and New York. (We ignore Texas because it has not been operative for very long.) PJM, New York and New England have had capacity obligations for some time. California has not.

In these few years, California has proved to be a spectacular failure, while the Northeast markets have quietly worked quite well. The outcomes in California were those that capacity obligations (or better, demand response) are designed to prevent: rolling blackouts and politically unacceptable price spikes. Meanwhile, the Northeast markets – inclusive of capacity obligations – have attracted seemingly sufficient amounts of investment and have avoided the egregious price spikes. There are so many factors involved that it cannot be concluded that the California situation could have been avoided with a capacity obligation, but on the other hand there is no evidence that the Northeast capacity obligations are not achieving exactly what they set out to do. The PJM market, for example, saw its capacity reserve margin become a binding constraint in late 2000/early 2001, capacity prices consequently rose and seemingly helped ensure sufficient resources were available for the following summer.

4.1. PJM

4.1.1. History

PJM has operated with perhaps the fewest problems of any of the regional electricity markets in the United States. It was established as an organization in 1927, and began operating the first fully functional ISO in the United States on 1 January 1998. PJM is both an independent system operator (ISO) and independent market operator (IMO). It includes all or part of 5 states in the Mid-Atlantic region, and the District of Columbia. It has recently expanded in the west ("PJM West" – comprising much of western Pennsylvania and West Virginia) and is likely to expand again in the near future as a result of FERC's Regional Transmission Organization (RTO) initiative – perhaps as far as to Chicago. In its current configuration, PJM has an electricity system similar in size to that of England and Wales (with a higher peak demand), although it serves only a population only half as large:

- 25.1 million people in its control areas
- 614 generation sources
- 62,445 megawatts of peak demand
- 298,011 gigawatt-hours of annual energy
- 13,100 miles of transmission lines
- 67,0000 megawatts of generation capacity

- Nearly 200 participants in its markets
- Connections at 32 locations on the high voltage system to neighbouring systems

PJM always placed an installed capacity obligation on each of its member utilities, and in June 1999 the obligation was modified with a new "Reliability Assurance Agreement" (RAA) being put in place. The RAA is a multi-party agreement and its stated purpose is "to ensure that adequate Capacity Resources will be planned and made available to provide reliable service to loads within the PJM Control Area, to assist other Parties during Emergencies and to coordinate planning of Capacity Resources consistent with the Reliability Principles and Standards".

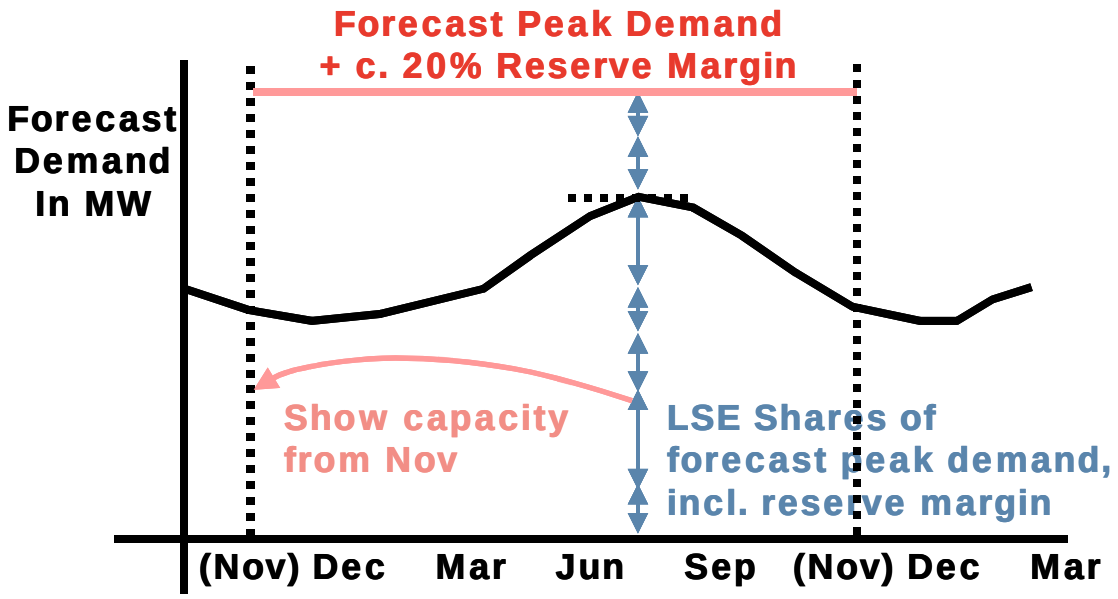
In the RAA, the historical capacity obligation, designed for vertically integrated utilities, was redefined in terms of "unforced capacity" and applies to Load Serving Entities (LSEs), which are not necessarily vertically integrated utilities. Unforced capacity is defined as installed capacity rated at summer (peak) conditions, down-rated for expected forced outages, and calculated for each plant on a rolling 12-month average, without regard to the ownership of contractual rights to the capacity of the plant. Capacity that is credited as unforced capacity undertakes to be recallable (ie, to curtail an export) and to supply its output into the PJM market as part of emergency procedures. To qualify to supply, it must show that its energy is deliverable to PJM load. It also undertakes to participate in maintenance coordination with PJM.

The PJM started to run markets for capacity credits in October 1998 (the first auction traded capacity for January 1999) and the product traded in this market was switched over to the new definition of "unforced capacity" in June 1999 to comply with the introduction of the RAA. The rules of the capacity credit markets are specified in the PJM Operating Agreement.

4.1.2. Definition of the capacity obligation

The starting point for the PJM's capacity obligation is a forecast that the pool operator compiles in November each year of peak demand in the subsequent summer. The sum of this peak demand and a reserve margin (of about 20%) is divided among LSEs in proportion to their customers' contribution to the forecast peak demand of the PJM system. (This allows customers to benefit if their peaks do not coincide with those of the system.) The peaks are evaluated in relation to the five hours of highest demand on the PJM system, which normally occur in July and August (although some areas and customers peak at other times.) Each LSE then has to show in advance (ie on every day of the subsequent year) that it has sufficient capacity to meet its forecast peak demand plus its share of the reserve margin. Figure 4.1 shows this process in diagrammatical terms, from the demand forecast, through the allocation to LSE's and back to the demonstration of capacity.

Figure 4.1
PJM Capacity Obligations



This scheme therefore relies on LSE's having capacity a short time (6-7 months) in advance of the peak period. If customers switch their retailer during the period after November, the associated capacity obligation transfers from the old to the new retailer, along with the customer. The PJM organises monthly and daily markets in capacity, to allow retailers to adjust their holdings and to bring them in line with their obligations.

Inside the PJM area, total generation capacity exceeds peak demand by about 7%. LSEs therefore have to buy capacity from generators in neighbouring control areas to meet the 20% reserve margin requirement. The PJM rules require evidence of firm transmission capacity up to and into PJM, and define the extent to which capacity imported from other control areas will count towards meeting a capacity obligation. These rules adjust the level of the capacity credit to allow for the reduction in reliability associated with use of transmission lines under the operational control of a third party, and for the reduction in reliability associated with a lack of short-term operational control over generators outside PJM. (Similar rules would be required for any capacity offered, for instance, by a French generator using the interconnector to the UK, or even by a Scottish generator, should separate systems apply in each area.)

The PJM rules also accommodate the possibility of interruptible load, in two different ways. If consumers shed demand voluntarily at peak times in response to prices or instructions from their suppliers (or if they switch on on-site generators that are not counted as capacity), they will reduce their supplier's share of the system's peak demand, and hence of its capacity obligation. In addition, if a consumer is prepared to grant control over its demand

management facilities to the PJM system operator, it is granted an "Active Load Management (ALM) credit", which also offsets its supplier's capacity obligation. This ALM credit is defined as the megawatt value of the consumer's interruptible demand adjusted by a special "PJM ALM Factor". The rules say that "the PJM ALM Factor will consider the reliability of the active load management, the number of interruptions, and the total amount of active load management", with the exact procedure being set out in a PJM procedural manual.

4.1.3. Deficiency charges

Daily capacity deficiency charges in PJM are based on the annual carrying charges of a new "combustion turbine" (better known in the UK as an "open cycle gas turbine"), installed and connected to the transmission system. This charge is then converted to an available capacity basis. Currently the daily deficiency charge is $176.83/MW. Paying this charge will raise the cost of a MWh on an average day by about $12.28/MWh.[9] If a LSE paid this charge every day for a year (365 days), it would pay $65,000/MW, equivalent to $65/kW-year or about twice the $30/kW-year cost of a peaking generator. This charge is a penalty rate, intended to encourage LSEs to build capacity and hence to reduce their total costs. However, if the LSE incurs the penalty for less than 170 days per year, it works out to be cheaper than building a peaking generator. Deficiency payments are allocated among PJM members, although the method of doing so has changed recently. (See section 4.1.7.)

4.1.4. Capacity credit market

The PJM capacity credit market consists of daily and monthly markets, in which market participants can buy and sell capacity credits through a process that establishes a market-clearing price.

- Daily market operation: The daily market is a day-ahead market. The daily market is useful so that LSEs can update and fine-tune their capacity positions on a daily basis, as retail loads are won and lost, and capacity contracts come on-stream or off-stream. The daily market is conducted based on the position of a participant for the market day estimated on the day the market is run. If a participant has a deficient position, PJM will only accept buy bids up to the deficiency amount. If a participant has an excess position, PJM will only accept sell offers up to the excess amount. Buy bids or sell offers are accepted into the daily market in order of priority. PJM strives to clear the market and post market results by 12:00 P.M. on the day the market is run.

- Monthly market operation: A monthly market may cover a period of one month or multiple months. The monthly market is a voluntary market where LSEs can plan ahead to match their capacity obligations with capacity credits.

---

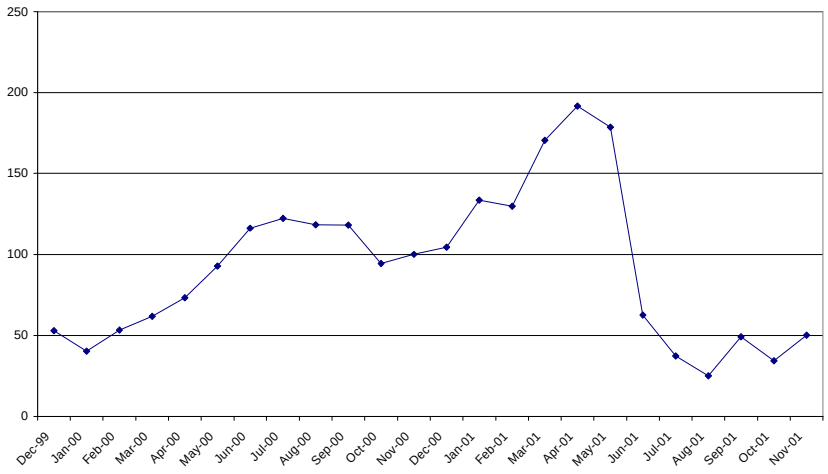
[9] This assumes that an average day has a system load factor of around 0.6. This also allocates the cost across all hours of the day, not just those hours in shortage. Hence, $12.28/MWh = $176.83/MW-day divided by 0.6 load factor and by 24 hours per day.

4.1.5. Capacity market outcomes

Since 1999, most capacity has been procured by LSEs through long-term bilateral contracts transacted outside of the PJM-run markets, or by virtue of the fact that the (remaining) vertically integrated utilities in PJM can use their own generation to offset their capacity for serving native load. Average MW volumes in monthly (or multi-month) auctions have been low – just under 200 MW. Average MW volumes in daily markets have been just under 800 MW. Generally it has been the smaller companies that have purchased a disproportionate amount of capacity through these auctions.

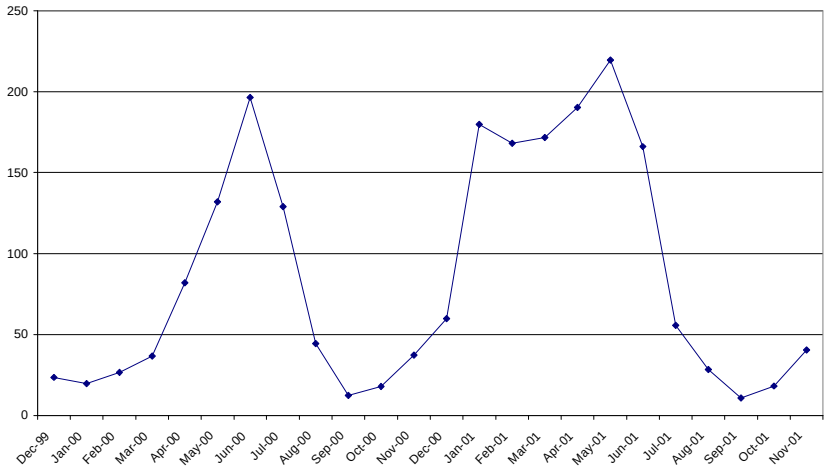
Prices in the PJM-run markets have been quite volatile, reflecting the deterministic nature of the capacity requirement and the way that the capacity obligation is set. Predictably, forward capacity prices have been the most stable because there is greater opportunity for entry to dampen price spikes, and for the effect of isolated events to be averaged in. The following chart shows ***12-month*** forward capacity prices in the years 2000 and 2001. Initially, PJM had sufficient capacity – prices were low, and were influenced by the opportunity cost of the right of recall held by PJM that could restrict capacity sellers from exporting to the east. Prices climbed during the summer of 2000, when a number of units went off-line, before settling back in the winter as it became apparent that the situation was temporary. At the end of 2000, PJM released the new capacity requirement figure for 2001, and it became clear that without additional resources, the system would only just meet the 2001 reserve requirement. The 12-month forward capacity prices quickly rose to the price of entry (which in turn equals the capacity deficiency penalty), before settling back in late 2001 as more resources entered the market.

PJM 12 Month Average Forward Capacity Price
2000-2001($/MW-day)



Three month forward capacity prices show a very similar pattern. The 3-month prices are more volatile because there is less opportunity for entry to mitigate a price spike, and short-term events are more fully reflected in the short-term prices.

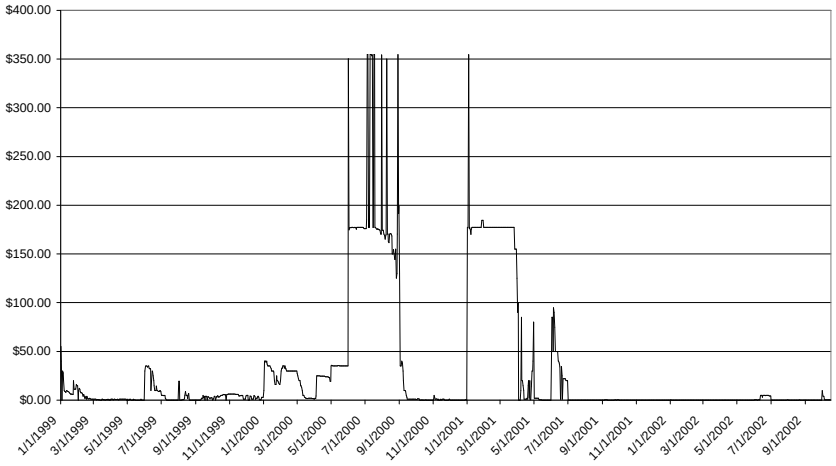
PJM 3 Month Average Forward Capacity Price
2000-2001 ($/MW-day)



In 2002, prices have been under $50/MW-day for most of the year, and fell to close as the end of the year approached and it appeared that sufficient resources are available.

Finally, the ***daily*** capacity market prices illustrate the bipolar nature of the capacity credit product. In days when the market has sufficient capacity, the daily price is zero or close to it. In days when there is a deficiency, the price rises to the deficiency level – again, in the summer of 2000, and again at the start of 2001 at the time the aggregate capacity requirement was increased. (Note: there are two deficiency levels: $176.83 when an LSE is deficient, and double that – i.e. $353.66 – when the market as a whole is deficient.)

PJM 3 Daily Capacity Price
2000-2001 ($/MW-day)



This volatility means that Load Serving Entities that have not covered their demand either by building capacity or by signing long-term contracts are exposed to substantial short-term risks. The market therefore tends to encourage vertical integration between retailers and

generators, either by retailers owning generation or through contracts. The ability of any company to cover its position (if it does not already own sufficient generation capacity) depends upon the liquidity of capacity markets, which has proven problematic.

4.1.6. Problems experienced with the capacity obligation system

One problem experienced by PJM was that some sellers of capacity had incentives to delist their capacity – i.e. renege on their capacity obligation - at short notice. Depending on market conditions in PJM and neighbouring markets, it could be advantageous for PJM capacity providers to commit to meet loads outside PJM even during times of capacity shortages within PJM.[10] Capacity can be diverted in this manner by delisting a capacity resource with two days notice.

If neighbouring systems such as the East Central Area Reliability Coordination Agreement area (ECAR) have prices spiking much higher than the prices in PJM, generators were tempted to divert capacity. As a result, PJM generators may be paid twice for the same capacity – once through capacity payments received during most of the year, and a second time from price spikes in neighbouring systems. The PJM consumers who pay for capacity will not receive the capacity precisely when it is most needed: i.e. when supply is tight and prices are high. Generators who delist from PJM in times of shortage are subject to penalties, including an installed capacity ("ICAP") deficiency charge and an increase in the outage rate assumed when calculating their unforced capacity quantity (which reduces their "unforced capacity"). However these incentives were widely recognized to be insufficient. With a deficiency penalty of $177/MW-day, a seller of capacity had to see an energy price differential of little more than $10/MWh on a standard 16 hour peak contract to be better off selling his power elsewhere.

Another perceived problem in PJM related to the volatility of the daily market and the way deficiency penalties were allocated. The rules in PJM allocate any deficiency revenues to entities that are long capacity after the day-ahead market has cleared. PJM alleged this led to some capacity providers having incentives to withhold capacity from the day-ahead market.

4.1.7. Recent changes

PJM changed the rules for evaluating compliance, so that compliance is measured on a seasonal basis. A much larger penalty applies if a LSE is short for a season, and so the consequence of delisting during a shortage are now much more punishing. The daily market still exists so that LSEs match changes in retail load on a daily basis.

---

[10] Refer to Hobbs, Inon, and Stoft, ***Installed Capacity Requirements and Price Caps: Oil on the Water, or Fuel on the Fire?,*** Electricity Journal July 2001

In response to the concern about capacity withholding by entities with surplus capacity in the day-ahead market, PJM changed the rules so that allocation of revenues from deficiency penalties is spread over all MW of all compliant LSEs, rather than just over the MW of the ones that are long. It is questionable whether this was a necessary response, but in any event, there has been an excess of capacity since the time when this change was made, and consequently daily capacity prices have remained close to zero.

4.2. New England

The New England Power Pool ("NEPOOL") began with two capacity markets, the monthly Installed Capability (ICAP) market and the daily Operable Capability (OPCAP) market. Both ICAP and OPCAP are structured in the same way. The NEPOOL ICAP market was implemented in April 1998.

4.2.1. Definition of the capacity obligation

NEPOOL operated a capacity obligation in real time, which is almost identical in operation to a real time capacity market. The pool operator compared each LSE's peak demand in each month with that LSE's rights to installed capacity. If the difference was too small to provide the required reserve margin, the pool operator imposed a deficiency penalty on the shortfall.

Such a system, which does not use a demand forecast, means that LSEs really need only fall short of their obligations at times when there is an actual system-wide shortage of capacity. As a result, capacity either has zero value (because it is in excess supply) or a shortage value (because there is currently a shortage).

4.2.2. Capacity markets

The ICAP market was a residual market. The difference between a participant's installed capability resources and its installed capability obligation (peak load plus installed operating reserves) was traded through an ISO auction.

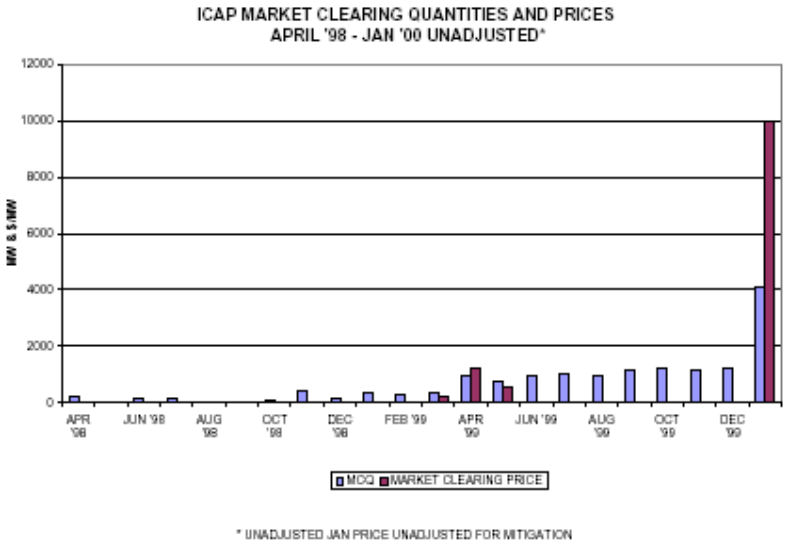
Trading in the ICAP market occurred every month. Bids were submitted in $/MW on the last day before the start of the month. At the end of the month, NEPOOL calculated a clearing price based on the bids of those participants with excess installed capability. The quantity traded was the sum of the quantities for those participants with a deficit. Participants who were deficient in installed capability paid the clearing price for each MW to those who were in surplus and who bid a price less than or equal to the clearing price. In 1998, FERC imposed a price cap of $8750/MW-month, pending the operation of the other markets (ancillary services).

Probably the most notable feature in New England was that the excess beyond serving obligations was based on actual monthly peak load and so could only be determined in real

time after the month-end. Aggregate deficiencies were matched against the bid stack of excesses, and the ICAP price was determined by the bid price of the last block of excess required to meet the aggregate deficiency. The price calculation generally occurred more than 30 days after the month in question based on the availability of meter readings. The amount cleared in the ISO-administered auction each month was generally around 1,000 megawatts prior to January 2000.

The average ICAP price from May 1999 to April 2000 was $3.35 and represented 8.7% of the total wholesale energy price during the same period.

During the June 1999 – January 2000 period the highest bid in each month increased from near $1,000/MW to $99,999/MW for significant blocks of ICAP. For most of the period a large quantity was bid at $0/MW and a small amount bid at high prices. The supply curve shape changed in January with significant quantities bid in the market at the highest price offered. This shift was the basis for the bid mitigation action taken by the ISO for the January 2000 market, which swiftly culminated in the ICAP market being terminated.



Source: ISO-NE

4.2.3. When and why was the ICAP market terminated?

After a study, the ISO concluded that the Installed Capability market was structurally flawed, did not send effective price signals for the construction or retention of needed capacity and was subject to manipulation. The ISO recommended terminating the market on 1 June 2000. NEPOOL retroactively reset ICAP market prices for January through March to zero.

FERC terminated the ICAP market effective 1 August 2000 and replaced it with an administratively determined deficiency charge. ISO-NE proposed a deficiency charge of $0.17/kW-month, but FERC rejected the ISO's proposed rate. In lieu it reinstated a ten-year

old administratively determined deficiency charge of $8.75/kW-month retroactive to August 2000. For the next few months there was an on-going controversy about the proper ICAP deficiency charge.

The termination of the ICAP market did not, however, lead to any weakening of the capacity obligation that lay behind it.

4.2.4. Current Requirements

The ICAP market was terminated last year. Load Serving Entities (LSEs) are still required to purchase sufficient ICAP each month to meet its allocation of the total NEPOOL ICAP requirement (ie, demand plus reserve margin), but participants who are short of capacity can only purchase more in the bilateral market or face the administratively-determined sanctions for failure to meet the existing ICAP requirement ("deficiency charge").

The current ICAP deficiency charge is $4.87/kW-month. Deficiency payments are allocated to suppliers with excess capacity and non-deficient LSEs.

4.2.5. Future capacity market

ISO-NE plans to implement a New-York style capacity market some time in spring 2003. In summary this market will have:

- A monthly ICAP Supply auction and deficiency auctions
- As in New York and PJM, net capacity adjustments for forced outages (based on a 12 month rolling average)
- Forward definition of the capacity obligation with full accounting for subsequent "load shifts" (customer switching).

The most notable change is perhaps the adoption of a ***forward*** obligation, defined some time in advance with respect to a demand forecast, as in New York and PJM. This proposal reflects the recognition that a capacity obligation enforced in real time or ex post exhibits price volatility that is little different from that seen in a pure energy market. It remains to be seen what form this new market will take, and whether it will make any difference to the pattern and level of capacity pricing.

4.3. New York

The arrangements in New York are very similar to those in the PJM. The NYISO's ICAP market evolved from the New York Power Pool's (NYPP), installed capacity requirement which NYPP has had since its inception. NY ICAP market started in November 1999.

4.3.1. Definition of the capacity obligation

The NYISO ICAP requirement (capacity obligation) is defined for a period called the capability year, which runs from 1 May in one year to 30 April of the following year. The capability year is further broken down into two capability periods, i.e., summer and winter. The summer period runs from May to October and the winter period runs from November to April. The capacity obligation applies separately to the summer and winter capability periods. (This seasonal definition allows obligations to be defined for summer peaking and winter peaking control areas, and adjusts the amount of capacity on offer for seasonal changes in generator availability.) As in the PJM, LSEs have to show in advance that they possess sufficient capacity to meet the installed capacity requirement.

In New York, the installed capacity requirement equals the NYISO peak load forecast for the next year, plus a percentage installed reserve margin determined by the New York State Reliability Council (NYSRC). An LSE must provide for 118% of projected needs, or go into the NYISO auction, or pay a deficiency charge. The deficiency charge is meant to provide a reasonable incentive to incur the cost of building a peaking generator. It acts as a de facto cap on capacity market prices.

ICAP obligations are defined for three zones: New York City; Long Island; and the rest of New York state. The details of these obligations vary by location, e.g. an LSE serving load in New York City must procure 85 percent of its ICAP from in-city generators.

In New York, ICAP providers are obliged to bid into the NYISO Day-Ahead energy market during the whole period of the capacity obligation.

4.3.2. Capacity markets

The NY ICAP market initially traded capacity for a six-month period. Last year, a monthly auction was added. The monthly auctions were designed to allow LSEs to bid to purchase or sell Installed Capacity to adjust for load-shifting (customer switching) during the prior month. Both type of auctions are divided to cater separately for the three zones (New York City, Long Island and the rest of New York state).

5. PROSPECTS

On 31 July 2002, FERC introduced a landmark proposal intended to bring more effective wholesale competitive to the bulk power system.[11] According to FERC's explanation in the Notice of Proposed Rulemaking (NOPR), the proposal for a Standard Market Design (SMD) is designed to create: genuine wholesale competition; an efficient transmission system; proper pricing signals for investment in transmission, generation facilities and demand reduction; and more customer options. Capacity adequacy proposals are also a key feature. The intention is that the standard market design will apply to all of the United States, based on best practices, with a minimum of required regional variations.

However, the NOPR is the very first step of a very long process. It appears that the resource adequacy provisions of the SMD NOPR, perhaps more than anything else, need further work. The intention is to provide for sufficient supply (generation) and demand (load management) resources to avert shortages. The main stated reason for the provision is that spot market prices alone are not expected to signal the need to begin development of new resources in time to avert a shortage, particularly since spot market prices are subject to market power mitigation measures and might not produce prices high enough when situations of scarcity arise. As FERC admits (at paragraph 113), "Because market power mitigation of spot market prices will tend to suppress the price signals for new entry, we are also proposing a non-price mechanism to assure that load meets a long-term resource adequacy requirement."

5.1.1. FERC's Proposal

FERC proposes that an Independent Transmission Provider must forecast the future demand for its area, facilitate determination of an adequate level of future regional resources by a Regional State Advisory Committee, and assign each load-serving entity in its area a share of the needed future resources based on the ratio of its load to the regional load.

In the proposal, the Independent Transmission Provider must assure that each load-serving entity in its area acts to meet its share of the future regional needs—through self-supply, contracts to purchase generation, biddable demand or other demand response program. The Independent Transmission Provider must apply standards, discussed below, to audit the adequacy of the plans of load-serving entities to meet the future resource needs of its area. Moreover, the Independent Transmission Provider must check that resources are not double-counted by different load-serving entities. In a region with more than one Independent Transmission Provider, each Independent Transmission Provider must coordinate this checking responsibility with all the Independent Transmission Providers in the region.

[11] 18 CFR Part 35, Federal Energy Regulatory Commission, Docket No. RM01-12-000.

If a power shortage occurs during which the Independent Transmission Provider is unable to satisfy demand in the spot market and also meet its reliability requirement for a minimum level of operating reserves, the Independent Transmission Provider must add a per-megawatt-hour penalty during the shortage to the price of energy taken from the spot market by a load-serving entity that did not meet its share of the regional needs for that year.

Further, if the operating reserve level decreases to the point that the Independent Transmission Provider must curtail load, the Independent Transmission Provider must, to the extent possible, curtail the spot energy purchases of the load-serving entity that did not meet its resource adequacy requirement before curtailing the spot energy purchases of load-serving entities that did. The load-serving entity is subject to such first curtailment during a shortage only in the amount by which it falls short of meeting its share of the resource adequacy requirement for the year in which the shortage occurs.

If a shortage remains after all such first curtailments are completed and additional curtailment is necessary, the remaining loads of the first-curtailed load-serving entities and the loads of other load-serving entities that have satisfied their resource adequacy requirement would be curtailed under the same protocol. In this case the shortage may be attributable to certain load-serving entities of either type that, whether or not they may have met their resource adequacy requirement. We expect that those load-serving entities that are short of their own reserves would lose service ahead of those that are not short, if selective forced outages are technically feasible.

5.2. Intentions of the Proposal

FERC's approach to resource adequacy proposed is intended to assure the development of both new supply and demand response resources. This approach is intended to focus on encouraging payment to fund construction of future resources instead of avoiding payment of a penalty for inadequate current resources as in some current programs.

FERC says that its proposal is intended to complement, not replace, existing state resource adequacy programs. A vertically integrated utility satisfying a current state resource requirement that equals or exceeds its share of the resource adequacy requirement would not have to do anything more. For those states that have retail choice programs in which retail customers or their suppliers buy power from a multi-state region, FERC intends this approach to provide for regional adequacy in a way that no one state alone may be able to accomplish. FERC is therefore acutely aware of the need to accommodate "load-shifting" (customer switching).

The intention of a ***forward-looking*** resource adequacy requirement is to create a demand for new resource entry in advance of a shortage, so that enough supply construction and demand response infrastructure installation are begun in time to avert the shortage. The planning horizon for each region is the number of years ahead for which the Independent

Transmission Provider must forecast annually its area load, as well as the number of years ahead for which load-serving entities must show that they have adequate resources. For example, the Independent Transmission Provider could forecast its area peak load three years from the present and require that each load-serving entity in its area have acceptable plans today to have enough resources three years from now to meet the forecast peak with a reserve margin of 12 percent.

FERC proposes to apply the requirements to all regions. In some multi-state regions, FERC has already approved an overall capacity requirement program that replaced individual state requirements. FERC's new proposals would replace these the current federally approved programs.

5.3. Comment on FERC's Proposal

The proposed capacity adequacy mechanisms are loosely based on the methods used in the Northeast. However, the proposed methodologies appear to suffer from a number of problems and it is not clear that they offer any improvement on the current installed capacity obligations used in the Northeast. For example, the reason for an LSE to comply with its capacity obligation is the threat of a penalty if it does not. The proposed penalty in the NOPR is added to the spot market price for the capacity-short LSEs when the system cannot meet the required level of operating reserves. If LSEs are free to contract at any time to avoid exposure to the spot market, they would not pay the spot market price and would avoid the penalty. It is difficult to see how the proposed mechanism could have its desired effect, if capacity-short LSEs could avoid the penalty by making sure they have energy in the NOPR. FERC is continuing to work on refining the resource adequacy provisions contract cover at peak times. It is also difficult to see how the size of the penalty could be commensurate with the value of capacity. FERC suggests penalties of the order of $500/MWh when operating reserves are violated. This seems too low, given recent estimates of the value of lost load and the price spikes needed to finance investment in generation.[12]

Furthermore, the method proposes that spot market service to a capacity-short LSE should be curtailed first, when a shortage is severe enough to require the shedding of some consumers. It is not clear that this rule is workable with existing technology, particularly in a region with retail competition where a retailer's customers may be widely disbursed within several distribution networks.

In any event, it is likely that the final market rules will be very different from the initial proposal contained, and the process of rulemaking has just begun.

---

[12] See Kahn Dr. A. E., "The Adequacy of Prospective Returns on Generation Investments Under Price Control Mechanisms", ***Electricity Journal***, March 2002.

6. CONCLUSIONS

6.1. Source of the Problem

The theory of competitive electricity markets predicts that prices will have to rise to substantially higher levels for a few hours per year on average, in order to provide enough revenue to cover the costs of building peaking capacity – generation capacity that only runs at times when demand reached the highest levels. Because such capacity only runs infrequently, it is efficient to build the type of plant that has low fixed costs and high operating costs. Other capacity within the portfolio of generators will run for greater numbers of hours on average; the more hours they are expected to run, the greater the cost saving to be made by adopting other technologies with higher fixed cost and lower operating costs. As a result, a competitive market paradigm predicts that investors seeking to minimise the total costs of meeting demand will build plant using a variety of technologies and (most likely) fuels. The price spikes needed to remunerate investment in peaking capacity are also needed to remunerate these other investments.

However, the market paradigm relies on the prospect of price spikes to encourage and reward investment in generator capacity. At times of peak demand, prices must rise to a multiple of electricity prices at other times – probably to several thousands of pounds per MWh, compared with today's electricity prices of around £20/MWh. These high price episodes are likely to be concentrated into one season, when capacity turns out to be inadequate because demand has risen suddenly and unpredictably and/or because some generation capacity is temporarily unavailable. In these conditions, when customers are being cut off, wholesale and (probably) retail prices are rising sharply, and some generators and traders are making large profits, the temptation for government bodies to intervene will be overwhelming. Knowing that this intervention will most likely lead to price caps (as observed in other markets), neither generators nor customers will be prepared to invest in sufficient generation capacity.

To create a favourable environment for long-term investment in generation capacity, it would be desirable to adopt a set of market rules that avoid the need for governments to cap electricity market prices in conditions of shortage. Capacity payments (as adopted under the old Electricity Pool) attempt to spread the cost of capacity over a wider set of periods. Capacity obligations attempt to impose the cost of meeting system capacity requirements directly onto electricity retailers – known in the US as Load Serving Entities. However, as the discussion above shows, only some capacity obligation schemes behave at all differently from pure electricity markets, and so only some capacity obligation schemes offer any hope of avoiding the problems identified above.

6.2. Review of US Experience

6.2.1. Prices, price caps and price volatility

Capacity obligations represent one method of ensuring adequate capacity in the face of inadequate returns in energy markets due to actual or threatened price caps. This is the strongest economic case for them, although there may be other benefits in terms of political accountability and the ability to set a higher level of reliability (for non-market reasons) than market economics alone would dictate.

However, many US capacity obligation schemes have not reduced short-term price volatility, but have merely transferred it into the capacity market. Such markets exhibit all the volatility of energy spot markets, if they are not subject to price caps (or low penalties for capacity deficiencies, which amount to the same thing).

The US schemes have not therefore been designed with the intention of reducing price volatility, but as a way to ensure sufficient capacity is built, as a carryover from the previous monopoly planning systems. However, ff both energy and capacity markets are subject to (binding) price caps, investment will suffer nonetheless from a lack of incentive.

US capacity obligations and markets do not even limit energy price volatility, since the obligation on owners of capacity is only to make the capacity available to short-term markets, not to limit its price. When capacity is short, prices rise in energy-only markets, regardless of the capacity obligations.

One way to avoid this problem would be to develop the obligation so that it does not apply to "capacity" separate from "energy", but instead obliges retailers to secure access to ***capacity*** (in the form of generation plant and contracts) in a way entitles them to purchase ***energy***. NETA already obliges retailers to sign contracts with generators in order to arrange output from half-hour to half-hour, but retailers are exposed to the wholesale price if these contracts do not cover their full load for an extended period into the future. Obliging retailers to cover ***all*** their forecast load with energy contracts lasting at least one or two years would prevent any retailer from being exposed to short-term price volatility. (Short term prices would still offer economic signals at the margin, but would have a limited impact on the financial position of retailers.) If they were not exposed to short-term prices, retailers would face fewer financial problems – and have less powerful arguments in support of price caps price caps - when and if a capacity shortage occurred.

Extending capacity obligations to cover energy entitlements would take them beyond the scope of any US schemes that we know. However, this extension would be necessary to give capacity obligations a clear role in avoiding short-term price volatility and its effects.

6.2.2. Effect on investment

The effects of capacity obligation schemes on investment are only as good as the incentives that they offer. If capacity deficiency penalties are set too low, they will act as a cap on capacity market prices and will discourage investment in just the same way as energy market price caps set too low. Even if a capacity obligation encourages investment, it cannot guarantee that price spikes will ***never*** occur. It can only promise that price spikes will be less frequent if (1) the scheme requires LSEs to build a greater volume of investment than the market would choose or (2) the scheme demands that LSEs expand capacity more smoothly (in line with demand forecasts) than they would otherwise (in response to expected prices), so that cycles of "boom and bust" are less pronounced. Unfortunately, there is simply too little experience of well designed capacity obligations, operating in a competitive market, to know if such effects are achievable.

6.2.3. Advance obligations

Recognising the undesirability of such price volatility, some US markets have concentrated on defining a capacity obligation by reference to a demand forecast, rather than by reference to actual demand. The PJM and NYPool systems have always operated on this basis, whereas the NEPool system is proposing to move over to it. The thought behind such a system is that the price of capacity needed to meet a forecast demand obligation will be more stable for two reasons:

1. In advance, the probability that capacity has a value lies somewhere between 0 and 1, and is more stable than in real-time (where capacity shortages occur with a probability of either 0 or 1). The probability weighted value of capacity is therefore also more stable;
2. By relating capacity obligations to a demand forecast, rather than actual demand, the obligation is more stable and less subject to fluctuations in real demand. As a result, capacity may have a value at times other than those of real shortage. (On the other hand, the scheme may fail to signal its value when a real shortage occurs but, as discussed above, the energy price will respond.)

Imposing a longer advance notice between the starting date of an obligation and the period to which it applies also gives LSEs more time to respond to signals by contracting for new capacity. However, LSEs can also react in advance to ***expected*** capacity shortages, when the capacity obligation is a short-term one. The adoption of a long advance notice period merely ensures that all LSEs see a signal in advance and receive a direct incentive to respond, rather than relying on their ability to forecast future signals.

6.2.4. Dealing with retail competition

Imposing capacity obligations in advance creates a need to re-assign obligations when customers switch from one retailer to another. Several systems in the US already achieve this reassignment, by linking the obligation to the customer (in much the same way as the

UK system links load profiles to a customer). These systems need to ensure that retailers continue to be held to their obligations (as defined by the original forecast), adjusted for customer switching ("load shift"). The examples we have examined organise monthly and daily capacity markets to allow retailers to trade their entitlements to capacity.

Capacity markets provide the flexibility needed to cope with retail competition, but only by exposing retailers to a certain amount of price risk. To the extent that they are gaining and losing customers, retailers have to enter the daily markets to buy and sell capacity, at which point they face a short-term price for capacity which may be highly volatile (as shown by the graphs in section 4.1.5. However, given the unpredictability of these prices, and of the delays in registering any customer switching, it does not seem likely that a retailer would be able to gain a competitive advantage by acquiring customers only when the price of capacity was low.

Incidentally, this residual exposure to short-term price risk is already present in energy markets (including those in the UK), because most wholesale contracts have a longer duration than retailers' contracts with their customers. Because of this mismatch between contract terms, the PJM is considering a proposal which avoids the need for retailers to make long-term commitments. Instead of sharing the capacity obligation among LSEs, it would be assigned in its entirety to the Independent System Operator (ISO) of the PJM system. (The equivalent in the UK would be the National Grid Company, or the "GB System Operator" to be appointed under BETTA.) The ISO would hold an auction to buy capacity from all possible sources in a least-cost manner and would then recover the costs of these purchases through a simple levy on retailers, applied to the demand at the time.

For instance, suppose that the ISO held an auction for "2005 capacity" during 2002. The ISO would estimate the capacity requirement and generators would submit offers to provide it. The ISO would match offers against the requirement and identify the market-clearing price. (We presume that potential new entrants would be allowed to participate, in order to widen the market, subject to their bearing heavy penalties if they fail to provide capacity by 2005.) The ISO would not pay this market-clearing price until 2005, at which time it would share the cost over all retailers, in proportion to their load at the time.

This scheme would entail even more centralisation of capacity decisions than a capacity obligation allocated to retailers (LSEs) and has not yet been adopted. However, its emergence as a possible candidate indicates the difficulty of reconciling long-term obligations on retailers with their inability (in law or in practice) to sign long-term contracts with all but the largest customers.

6.3. Design of a Possible Capacity Obligation Scheme

Any capacity obligation scheme consists of a number of rules covering the following points:

1. estimates of the ***future loads of all parties***
2. the ***frequency*** with which capacity adequacy is reviewed and the ***future period*** to which the obligation applies;
3. standards for ***required reserve margins;***
4. standards for the ***accounting of capacity***;
5. ***operational rules*** must be created to ensure that the capacity promised by generators is actually forthcoming when needed;
6. ***penalties*** for failure to meet the obligations.

Within the UK context, it is not difficult to imagine what these rules might say and how a capacity obligation scheme might be implemented.

NGC already possesses a method of forecasting long-term peak demand (item 1), for its Seven Year Statement, although these forecasts have so far had no direct commercial implications. NGC produced short-term demand forecasts for the Electricity Pool, and accepted a share of the costs due to its forecasting errors under the System Operator incentive scheme. A similar approach might be applied to forecasting peak demand a year or two ahead.

The process for allocating total peak demand among retailers could build on the existing billing software, which not only attributes every customer to a supplier, but also contains load profiles for particularly customer classes. These profiles could be used in sharing out the total peak demand forecast.

The reserve margin (item 3) is not defined at present, but does not seem to vary much between different systems. Figures just above and below 20% appear to be common. On the other hand, rules for defining capacity (item 4) are bound to be complex; however, examples already exist in the US (and in UK power contracts).

The obligation on generators to make capacity available (item 5) seems most likely to translate into an obligation to offer uncontracted capacity to the Balancing Mechanism (or else to incur a penalty). The size of the penalties (item 6) would require careful consideration, since they depend upon the nature of the scheme and likely values of the loss-of-load probability at the time when the obligation applies.

The time at which the obligation applies (item 2) is perhaps the most important item in all capacity obligation schemes, given that the purpose is to smooth prices. As shown by a comparison of PJM and NYPool with NEPool, a capacity obligation does not smooth prices

at all, if it is only enforced in real time, of ex post. In such conditions, as the experience of NEPool highlights, capacity has no value when it is in excess supply, and a value equal to the deficiency charge when there is an actual shortage. Such a scheme has no impact on the kind of price volatility that makes energy-only electricity markets vulnerable to political intervention.

US experience shows that long-term trading of capacity obligations defined for a future date provides more stable prices than short-term trading of capacity obligations that apply immediately. A long-term market would require a capacity obligation applying to peak demand some time (2 or 3 years) in the future, and would then charge a penalty for each day that a retailer was unable to demonstrate sufficient capacity to meet the obligation. This penalty would effectively spread the costs of capacity shortage over a wide period. (However, it would also create overlapping obligations if reapplied every year.)

However, such markets have to solve additional problems, if they are to be successful. In particular, in order to make a capacity obligation work in the context of retail competition, it is necessary to have some way to transfer the obligation. PJM has developed a system that seems to work. Each distributor is assigned to a zone. Each zone is assigned an annual capacity obligation by the ISO for the next year. The distributor assigns each customer an annual capacity obligation ***that stays with the customer for a year***. Retailers have to meet the capacity obligation on every day during the year, for all customers they are serving.

Hence, when the customer switches retail suppliers, the capacity obligation switches and the receiving supplier has a higher capacity obligation to meet thereafter on a daily basis. The PJM offers a daily capacity market for the purpose of trading capacity required or freed up due to switching, but such trading does not need to be organised centrally, as long as ownership of capacity obligations and capacity resources is duly registered for each retailer.

This description shows that there is a workable way to accommodate retail competition and a capacity requirement.

6.4. Costs and Risks

Imposing a capacity obligation on retailers does not necessarily avoid the kind of price spikes that make energy-only markets unsustainable. Capacity obligations and capacity markets are a prominent feature of several restructured electricity markets in the United States. Competitive wholesale electricity markets have worked reasonably well in the areas that have a capacity obligation, while the one area that lacked a capacity obligation, California, experienced a market collapse (bankruptcy of the Power Exchange and a utility). It is far from clear that capacity obligations and success are linked through cause and effect, however much has been made of the apparent link in public discussion.

The imposition of capacity obligations is predicated on the notion that an energy (kWh) market alone will not operate in the manner needed to encourage efficient investment

decisions. The aim of a capacity obligation system is to encourage investors to build capacity even though electricity market prices are or will be capped. The amount of capacity that results depends on the scale of the obligation and penalty for non-fulfilment. The effect on price spikes depends on the way in which the obligation is defined and enforced.

The centralised estimation of future loads and reserve margins undermines the ability of competitive market forces to determine the optimal level of capacity and, historically, has led to levels of capacity that probably exceed the optimal level. The costs of excess capacity brought about by a capacity obligation must be therefore compared with the cost of inadequate investment due to market failures.

NERA recently estimated the cost of excess reserve margins, as shown in Table 6.1. Here, rows (1) to (6) estimate the reserve requirement based on data about total consumption, a load factor (to convert consumption into peak demand) and a reserve margin obligation of 20%. Rows (7) to (10) calculate the annual cost of this reserve margin at an annual unit cost of capacity of $31 or £20 per kW and the ***excess*** cost of this capacity, assuming that the market would choose a reserve margin of only 8% (ie that 60% of the assumed reserve margin of 20% is excess). Rows (11) to (13) compare this figure with the total cost to consumers of their annual purchases of electricity. For the US, this estimate was about $2.5 billion per annum, or 1% of the retail bill. Given the UK's lower population and per capita consumption, the comparable figure for the UK is much lower – only about £150 million per annum or 1.1% of the retail bill.

Table 6.1: Cost of Excess Reserve Margin

Row	Source	Item	US Parameters	UK Pameters
(1)	IEA data	Total consumption	3500 TWh	330 TWh
(2)	(1) / 8760h p.a.	Average consumption per hour	400 GWh	38 GWh
(3)	NERA	Load factor	60% GWh/GW	60% GWh/GW
(4)	(2) / (3)	Peak demand	666 GW	63 GW
(5)	NERA	Reserve margin	20%	20%
(6)	(4) * (5)	Reserve capacity requirement	133 GW	13 GW
(7)	NERA	Annual unit cost of reserve capacity	31 $/kW-yr	20 £/kW-yr
(8)	(6) * (7)	Annual total cost of reserve capacity	4,129 $ million	251 £ million
(9)	NERA	Of which, excess portion =	60%	60%
(10)	(8) * (9)	Excess annual cost of reserve	2,477 $ million	151 £ million
(11)	NERA	Average retail price	72.5 $/MWh	40 £/MWh
(12)	(1) * (11)	Total annual bill of consumers	253750 $ million	13200 £ million
(13)	(10) / (12)	Excess annual cost of reserve (%)	1.0%	1.1%

Institution of a capacity obligation (with or without a capacity market) has no implications for diversity. All capacity usually counts the same towards the obligation and the energy market encourages the least-cost choice among different fuel types, etc, as per the energy market paradigm. If the capacity obligation demands investment in additional capacity that will not run often, cost pressures will tend to encourage the construction of additional peaking capacity – in the UK, usually an open cycle gas turbine with an annualised capital cost of about £20 per kW. However, other choices between different technologies and fuels, for capacity that is expected to run, will be unaffected.

6.5. Appraisal

6.5.1. Purpose of US capacity obligations

Capacity obligations systems in the US were not created to solve the problem of price spikes, but to decentralise capacity planning, which was formerly a component of the monopoly's obligation to serve. The allocation of capacity obligations to individual LSEs allows the market to determine the value of capacity for a particular reserve margin. In principle, the outcome of such a market process will be more efficient than the outcome produced by a centrally determined capacity payment formula. In fact, most schemes offer a mixture of capacity obligations and payments. To the extent that LSEs fail to sign contracts for capacity, they will pay a penalty and the revenue from such penalties is allocated to generators holding capacity. If no LSE signed any contract, the penalty would be akin to a capacity payment from the central pool; however, the obligation allows LSEs to seek out capacity on their own behalf and to determine its value in negotiations.

The scheme still requires a central authority to set the required level of capacity for the market as a whole. This level may be different from the level that an energy-only market is likely to pick. If they require more capacity, as is likely, capacity obligations will impose additional costs.

The retention of a centrally determined obligation derives principally from a fear that liberalised markets would not provide an efficient level of capacity, owing to the "shared" or "public good" nature of security of supply. The extra cost imposed by the capacity obligation would therefore be justified by the benefit of avoiding undesirable load shedding. In fact, imposing a high price on energy taken at short notice during a capacity shortage would impose the cost of failing to invest on individual market participants, thereby giving them an incentive to invest in sufficient capacity. However, energy prices in the US markets studied above are subject to price caps and potential investors might believe that price caps could emerge in the UK electricity market, if prices ever rose during a capacity shortage.

Such price caps (real or threatened) depress the incentive to invest in capacity in an energy-only market. The creation of a capacity market (along with an obligation on LSEs to buy capacity) provides an alternative source of revenues for investment in generation and an additional incentive to invest in generation capacity.

6.5.2. Capacity prices and investment incentives

In the US, the incentive to build capacity is only as strong as the penalty for failing to meet the capacity obligation. That penalty, which acts as a de facto cap on capacity prices, is set on a daily or other short-term basis. LSEs can trade capacity over short intervals, so the effect of the penalty depends on how often and when a LSE chooses to run short of capacity. Setting a low price cap on capacity prices discourages investment just as effectively as a cap on energy prices.

Capacity has a large and positive value when it is short, and zero value otherwise. Because of this "bi-polar" pricing, capacity obligations may not avoid troublesome price spikes. Indeed, the NEPool scheme showed that a capacity obligation defined with respect to ***current actual demand*** produces the same kind of price volatility as a short-term energy-only market. The PJM and NY Pools both define capacity obligations in relation to a ***demand forecast*** that remains constant over a whole year. As a result, capacity may appear short relative to the obligation (and hence price spikes may occur) at times other than when capacity is really short (in relation to actual demand). Basing the obligation on a forecast therefore has at least one advantage: the capacity price spikes (and apparent profits) need not coincide with the consumer hardship of forced load shedding due to actual capacity shortages. The resulting increase in revenues may therefore be less problematic for consumers, politicians and regulators, and less likely to provoke further intervention in pricing.

6.5.3. Effect on volatility

US schemes (being designed to circumvent energy price caps) separate the trading of "capacity" from the trading of "energy". In practice, this means that a generator selling capacity is only obliged to offer its plant to the system operator for short-term balancing, at any price up to the current price cap. Hence, the original capacity obligation schemes did not (and were not intended to) reduce the potential for overall (energy + capacity) price volatility, except to the extent that they call forth additional capacity and make real shortages less frequent. However, the volatility of capacity markets has recently been viewed as a problem, leading to the suspension of some capacity markets. The US is therefore in the process of examining proposals (chiefly a move to longer term obligations) designed to reduce volatility.

As long as capacity remains separated from energy, capacity obligations are unlikely to have much impact on the level of price spikes (even if they reduce their frequency). When a shortage occurs, generators are still entitled to offer their plant to short-term balancing markets at any price (up to a price cap). In order to or dampen this kind of volatility, it would be necessary to adapt the concept of capacity obligations, so that LSEs were required instead to show evidence that they had secured energy contracts (as well as "capacity") sufficient to cover their share of forecast demand and the reserve margin.

Under NETA, this kind of rule would look like an extension of the current obligation on market participants to match contracts to demand (and generation) an hour in advance. If market participants were required to match contracts to a forecast of peak demand from six or twelve months in advance, they would have to secure contract cover for each peak period. This would prevent any retail supplier from entering the peak period exposed to the risk of price spikes, so that they would have less reason to call for intervention when price spikes occurred (although consumers would still be exposed to retail price increases unless they had contracts of a similar duration).

6.5.4. Implementation requirements

The fact that capacity obligations apply in US electricity markets indicates their feasibility. They require additional rules on at least the following topics:

- Defining the demand forecast and the share allocated to each LSE (retail supplier);
- Defining what capacity generators (and other sources, such as imports) have to offer;
- Registering the capacity (or contract) holdings of each LSE;
- Setting a penalty for deficiencies; and
- Comparing holdings with obligations to determine the penalty charge for each LSE.

Implementing such a reform through the Balancing and Settlement Code (which enacts NETA)

6.5.5. Markets and competition

The schemes observed in the US, and the possible adaptation for the UK described above, do not diminish the efficiency-enhancing properties of competition, except to the extent that a central body determines a reserve margin (and each LSE's share of it) and defines what capacity each generator has to offer. The outcome of this centralised process may not be the efficient level of investment that would emerge from an energy-only market. However, if an energy-only market is subject to (or likely to invoke) price caps or other interventions, then the efficient outcome is unattainable and the central determination of capacity requirements reduces efficiency less - and may even increase it. The process of securing contracts for capacity would still allow buyers to seek out the least cost sources from the range on offer.

The US systems have adapted to cope with retail competition. They calculate each LSE's obligation as the sum of individual obligations attributable to its customers. When customers switch supplier, they carry their obligations with them. The new supplier is required to hold extra capacity, whilst the old supplier can hold less. This process would not affect the ability of different suppliers to serve customers.

6.6. Summary

The basic structure of a US-style capacity obligation makes electricity retailers proportionately responsible for ensuring that adequate spare capacity is built to meet their loads. In doing so, however, both theory and practice have demonstrated that only carefully designed systems are likely to achieve their objectives. In particular, experience seems to suggest that:

- only a scheme which defines capacity obligations some time in advance will avoid creating price spikes at precisely the same time as an energy-only market would;

- any scheme needs to tie capacity obligations to customers (so that they transfer when the customer switches from one retailer to another) and to enforce the obligation continuously (to avoid distorting competition between retailers and incentives to invest in capacity);

- the success of the capacity obligation scheme in promoting investment depends upon the ability of capacity market prices to rise sufficiently high, on enough occasions, to offset the loss of revenue caused by energy market price caps, but imposing low price caps (or low deficiency penalties) on capacity markets would undermine investment incentives just as easily.

Even in the US, where capacity obligation schemes are practised most widely, there is no experience of such a scheme that would lead one to conclude that they either promote more investment or achieve more stable prices than an energy-only market. US schemes have not been primarily motivated by a desire for price stability and some appear to be a method of circumventing energy market price caps (to secure an additional source of revenue).

Some of the design features needed to reduce exposure to price volatility (and with it the incentive for customers to lobby for government intervention) have not yet been tried out in the US, although some are under discussion. Even the latest proposals from FERC are only beginning to identify possible solutions and have not yet led to the creation of new and successful schemes. The UK would be leading the way in electricity market design if it pursued a capacity (or contracting) obligation designed to reduce overall price volatility.

APPENDIX A. ELECTRICITY MARKET PARADIGM

A.1. Generation Cost Conditions

The paradigm begins with the recognition that electricity is generated by a variety of technologies, each of which has different cost characteristics.[13] For clarity of exposition, the paradigm is normally presented in terms of three technologies as follows:

Table A.1
Cost Characteristics of Typical Production Technologies

Type	Annual Fixed Cost	Variable Cost per Unit
"Baseload"	High	Low
"Mid-Merit"	Mid	Mid
"Peaking"	Low	High

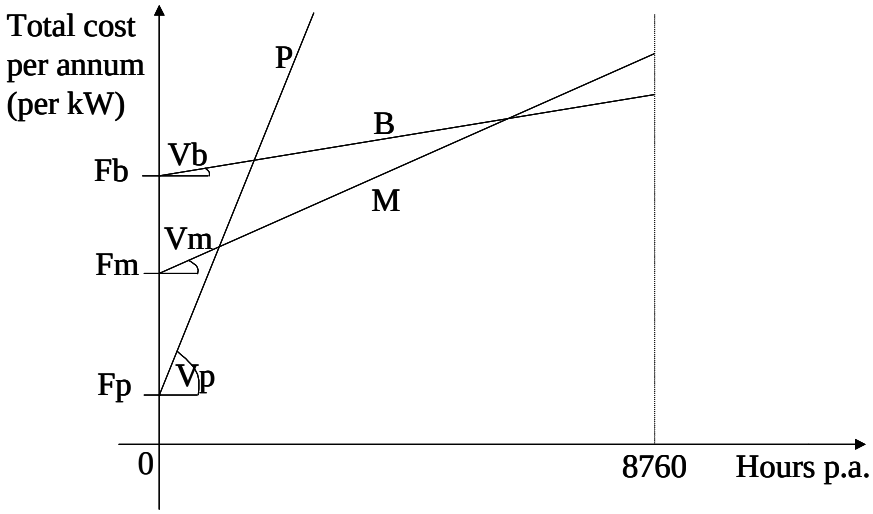
In practice, the nature of these technologies changes from time to time. In 1990, baseload generators included both "must-run" plant (run-of-river hydro and nuclear) and coal-fired generation; mid-merit plant meant oil-fired generation. By 2000, however, gas-fired generation had largely supplanted the role of coal as a baseload technology and coal-fired generation had been pushed into mid-merit. Recent rises in gas prices and the change in market arrangements have reversed this trend to some extent. These changes show the risks associated with long-term investment in generation and the associated pressures for risk management (for example, diversification of fuel sources). The paradigm takes costs as given and users must therefore make separate allowance for risk.

Figure A.1 shows the total annual costs (per kW of capacity) for the three basic types of generation: baseload (b), mid-merit (m) and peaking (p). Each type of plant is represented by an annual fixed cost per kW (F) and a variable cost per kWh (V). Total annual costs therefore vary according to the number of hours that capacity runs in a year, up to the maximum of 8760 hours (= 24 x 365). For example, a kW of baseload capacity has a relatively high annual fixed cost of Fb, which the plant incurs even if the plant does not run for any hours during the year. Hence, at 0 hours (on the left hand side of the figure), the annual cost per kW for baseload capacity is Fb. If the plant runs, its costs increase, at a rate equal to Vb, the unit cost of output from baseload capacity. The line marked B therefore represents the relationship between annual hours of operation and total annual costs for

[13] One may ask why cost conditions should fall into this neat pattern and the answer is "by elimination". Any plant for which both costs were high, or for which one cost was high and one middling, would be uneconomic compared with at least one other technology. As a result, it would be ignored or devalued, until either its fixed or variable costs fell sufficiently to make it competitive against the technologies shown, in which case it would replace one of them. In principle, costs need not be only "high", "middling" or "low", but can take any value. As a result, it is possible for many technologies to be economic. However, in practice, it is rare for more than three technologies to achieve least-cost status at any one time. Other technologies may be left over from previous eras but must compete with those that are currently least-cost.

baseload plant. (This report will refer to "hours" even though electricity markets may use half-hours or even shorter periods for settlement purposes. The analysis applies for different settlement periods, but the arithmetic is complicated by conversion factors.)

Figure A.1
Annual Costs per kW for Different Types of Generation



Mid-merit plant has a slightly lower annual fixed cost per kW of Fm, but a slightly higher variable cost per kWh of Vm. The cost function of mid-merit plant is given by the line marked M. At zero hours of operation, the cost of mid-merit plant is lower than the cost of baseload plant, but its costs rise rapidly if it runs. Mid-merit plant is more expensive than baseload plant if running for most of the year (ie, line M rises above line B).

Similar, for peaking plant, the annual fixed costs per kW is very low, at Fp. However, its variable cost per kWh, Vp, is very high, so that the total annual costs of peaking plant are very high, if it runs for a lot of hours in the year.

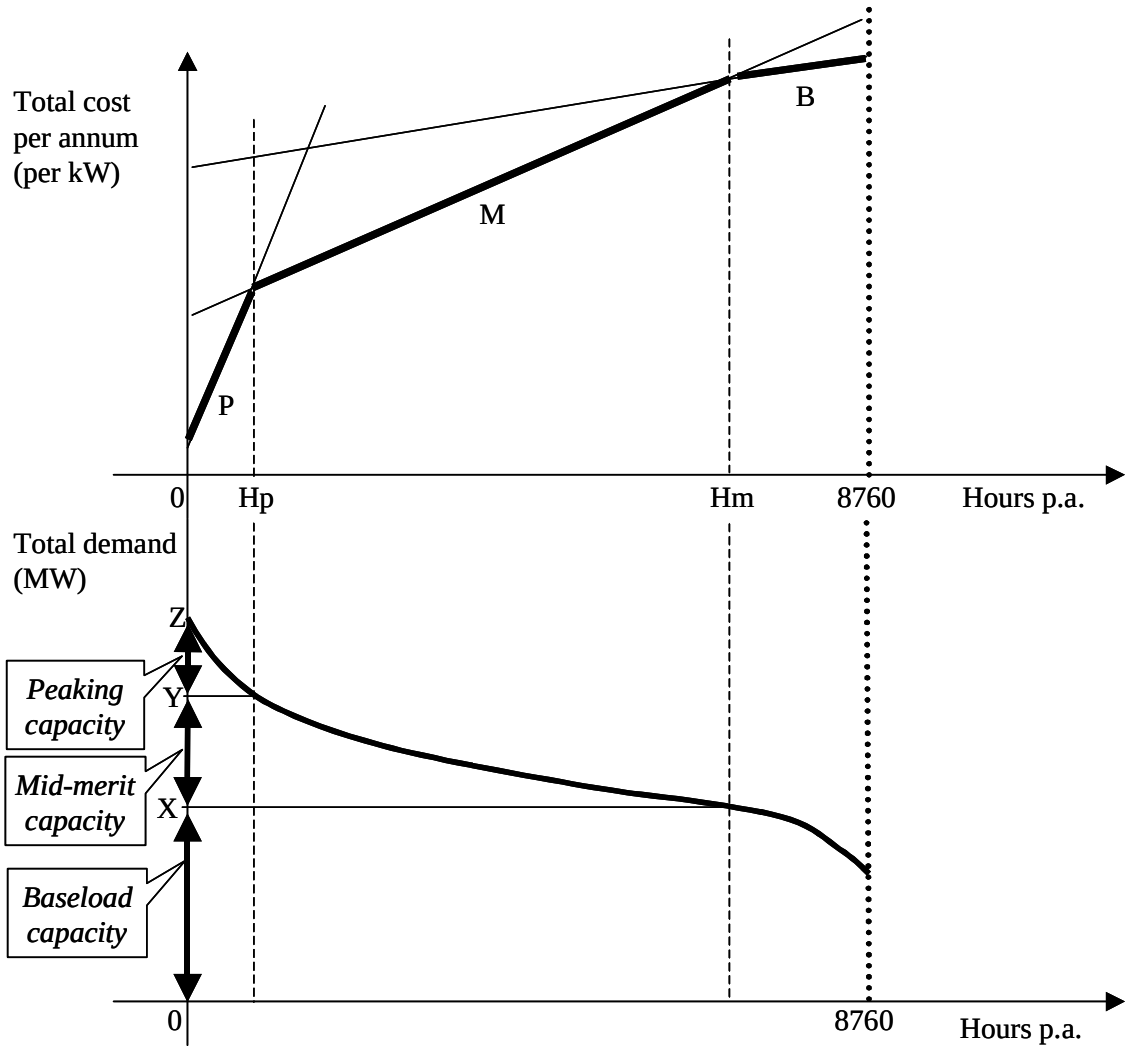
Given these three technologies with different cost conditions, it is possible to calculate an efficient pattern of investment in each one, sufficient to meet peak demand with the least total cost.

A.2. Efficient Least-Cost Investment

In planning terms, it is possible to identify an efficient portfolio of investment, by comparing the cost conditions in Figure A.1 with the demand conditions shown in Figure A.2. Here, the top half of the figure shows the same three cost functions, B, M and P, as in Figure A.1. The darker line indicates the least-cost technology at different levels of operation over the year. If the plant is expected to run for only a few hours (Hp), then peaking plant is the cheapest form of generation, because its low fixed costs outweigh its high running costs. For plant expected to run between Hp and Hm hours, mid-merit plant has the lowest costs

overall. For any plant expected to run more than Hm hours per year, baseload technology is the cheapest, as its low variable costs more than compensate for its high fixed cost.

Figure A.2
Efficient Capacity Planning Model



The lower half of the figure shows a "load duration curve", a conventional presentation device in the electricity industry. On the left hand side is the hour with highest demand; on the right hand side is the hour with lowest demand. Other hours are arranged between them. Reading horizontally, one can see the number of hours ("duration") in which demand is equal to or higher than any particular level. (For planning purposes, the measure is a forecast, based on historical data.)

The figure shows the boundaries Hp and Hm transposed onto the load duration curve. The interpretation of this graph is that a least-cost portfolio of generation would include baseload capacity of OX, mid-merit capacity of XY and peaking capacity of YZ. Then if, in

any hour, generation capacity operates in least-cost order ***with respect to its variable costs***, each plant will run for a number of hours per year at which it represents the least-cost technology. Baseload capacity will run for long periods (greater than Hm), ie nearly all of the time. Mid-merit plant will run for up to Hm hours, but will be restricted to hours when demand is higher than OX. Peaking plant will only run when demand exceeds OY, and will not run for more than Hp hours.

A.3. Energy Revenues, Capacity Payments and Security Standards

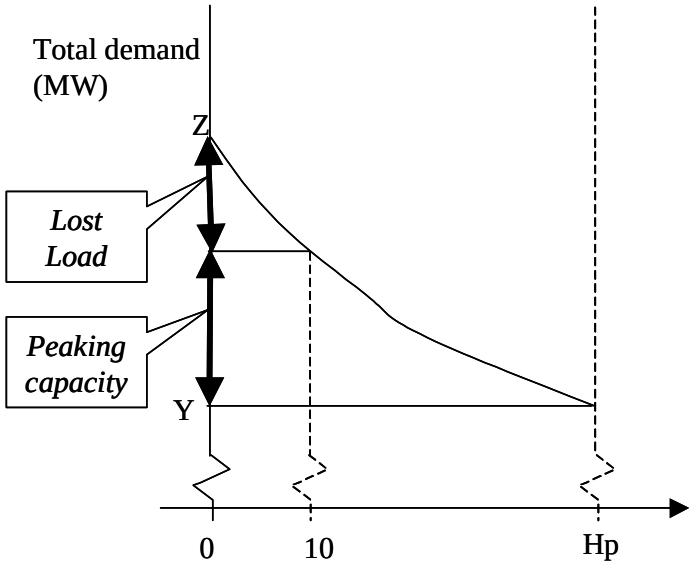
Even under monopoly systems, central planners noted the implications of this paradigm for tariffs and revenues. They assumed that the efficient (wholesale and retail) price per kWh of energy would be the "system marginal cost", ie the marginal cost of meeting extra unit of demand. In hours to the right of Hm, that cost is represented by Vb, the variable cost of additional output from baseload capacity. In hours between Hp and Hm, system marginal cost is Vm, the variable cost of mid-merit capacity. For hours below Hp, the system marginal cost would be Vp, the variable cost of peaking capacity, but for the following complications.

First, the capacity planners noted that, if prices or tariffs only reflected variable costs, they would not recover the whole costs of the system – which would prevent efficient investment from taking place. Further examination of the cost structure set out above proved that the "missing" amount of revenue was equal to Fp, the fixed annual cost of peaking generator, multiplied by the total amount of capacity on the system. This observation led to the design of various "capacity payments" to augment energy charges.

Second, however, planners noted that building generation capacity sufficient to meet demand of OZ was inefficient, since consumers were not willing to pay as much as Fp for consuming energy in the hour of peak demand. They usually adopted a security standard, for instance a policy of anticipating 10 or 20 hours per annum of "lost load". Figure A.3 is merely a close-up of the section of peak demand section in Figure A.2, but shows the rational efor such standards.

In this case, the planners have decided to build total capacity that is sufficient to meet total peak in demand in all but 10 hours. In those 10 hours, demand would rise higher than the level of total capacity, so some demand must be cut-off ("lost") by the system operator. In – principle, the efficient level of lost load depends on a trade-off between capacity costs (Fp) and the "value of lost load" (VOLL).

Figure A.3
Peak Demand Conditions



Suppose that the fixed cost of building and maintaining a kW of peaking capacity (Fp) is equal to £20 per year. (Assume also that variable costs, Vp, are relatively small and can be ignored here.) Suppose consumers are willing to pay £2/kWh to maintain supplies at times of peak demand. The planners have a choice:

1. Build one more kW of peaking capacity at a cost of £20 per year; or
2. Cut of 1 kW of load for 10 hours per year.

The second option means an extra 10 kWh of lost load per year, at a value of £2/kWh, giving a total cost of £20 per year, the same as the cost of building peaking capacity. If there were less peaking capacity, there would be more hours of lost load, and option 1 ("Build") would be cheaper than option 2 ("lose load"), so there would be a signal to invest. If there are less than 10 hours of lost load per year (on average, taking several years together), there would appear to be excess capacity, which provides a signal that it may be efficient to close plant. A similar set of comparisons between the costs of peaking, mid-merit and baseload stations will show what type of plant should be built, depending on the number of hours it is expected to run. In practice, the answer is usually a peaking plant or a baseload plant: mid-merit capacity is nearly always old plant displaced from a baseload role by newer, more efficient technologies.

The choice of security standard (i.e. the desired average annual number of hours of lost load) therefore determines the level of investment in capacity in a planned system. This approach is not directly relevant to a market system, but it can be adapted to show what an electricity market would look like, if it offered similar incentives for efficient, least-cost investment.

A.4. Extending the Market to Cover Market Conditions

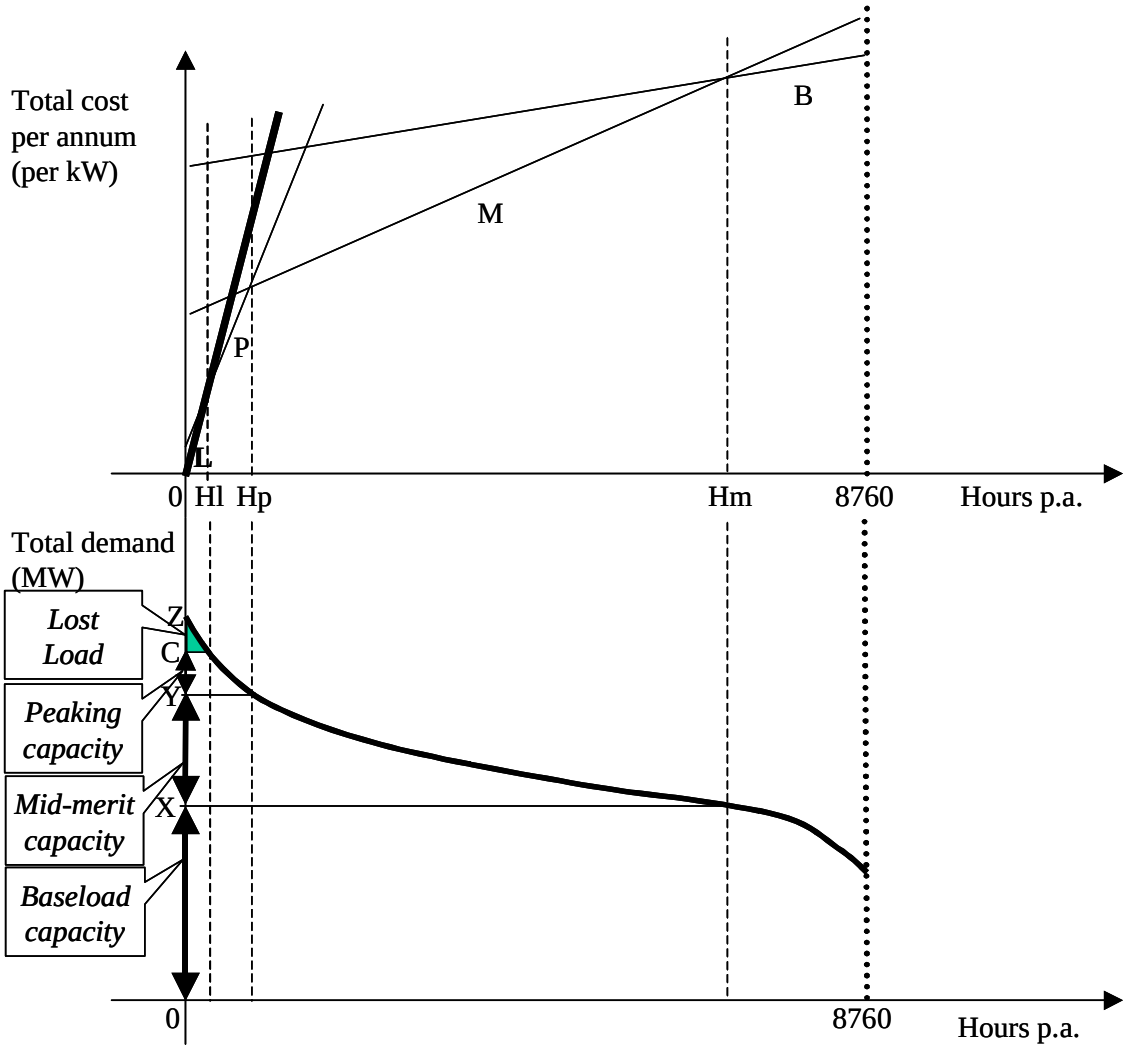
To convert the analysis above into a market paradigm, it is necessary to recognise that demand-response – in the form of load-shedding – is an alternative method of balancing supply and demand, to be placed alongside the generation technologies shown in Table A.1. For simplicity, we will examine only the possibility of losing load at peak times and will refer to a single value of lost load, meaning that value which applies at peak times, when consumers are being forcibly interrupted. Table A.1 can then be extended with a "fourth technology", losing load, which has a zero annual fixed cost, but a very high variable cost per kWh of lost load, as in Table A.3.

Table A.3
Extend Table of Cost Characteristics

Type	Annual Fixed Cost	Variable Cost per Unit
"Baseload"	High	Low
"Mid-Merit"	Mid	Mid
"Peaking"	Low	High
"Lost Load"	*zero*	*VOLL*

This information can be added quite simply to the diagram of cost conditions and capacity planning, by incorporating an additional line, L, to represent the possibility of losing load.

Figure A.4
Market Determination of Capacity Investment



This line creates another cross-over point, Hl, below which losing load is cheaper than building and using peaking capacity. The efficient decision is therefore only to build capacity up to the level now marked as "C" and to shed load when demand exceeds that level. The volume of lost load is shown in Figure A.4 as a shaded triangle in the top left-hand corner of the load duration curve.

A.5. Implications for Electricity Market Pricing

The implications of this adaptation become apparent when applied to a competitive electricity market, given two key assumptions about the way in which competitive markets work:[14]

1. Producers will select the least-cost combination of output. Given that fixed costs are unavoidable from hour to hour, producers will minimise variable costs by running only baseload plant when demand is low, calling on mid-merit plant only when demand rises above OX and using peaking plant only in the rare hours when demand exceeds OY. Producers will continue to serve demand as long as they have capacity available and load-shedding will only be necessary if demand would otherwise exceed OC.
2. Market prices will settle at the marginal cost of the most expensive producer chosen or at VOLL. Competition normally drives down prices to marginal costs and the electricity sector is no exemption from this rule, although the marginal costs for the market (or "system") will be determined by the plant type with the highest variable cost that is called upon in the least-cost combination.

As before, this means that the price of energy will equal the "system marginal cost", but in the hours of lost load there is a new definition of this term. Instead of peak prices being set equal to the variable cost of peaking plant, Vp, with some capacity payment required to recover total costs, prices are now set equal to VOLL – the variable cost of shedding load – in the few hours when demand exceeds capacity.

The revenue from sales at VOLL in these hours substitutes for the capacity payment described in section A.3 and, in an efficiently built system, will equal the capacity cost of a peak (as per the analysis in that section). If there are a lot of hours of lost load, these revenues (annual hours of lost load x value per kWh of lost load) will exceed the cost of adding capacity (annual cost per kW) and will encourage investment. If there are few hours of lost load, there is excess capacity and it may be efficient to close plant.

In an electricity market, therefore, the incentive to invest depends crucially upon the market price that applies when load is being lost, and the expected number of hours in which load will be lost over the life of the investment. Box A-1 provides an equivalent analysis in terms of conventional supply and demand diagrams.

[14] The decision to replace monopolies with markets rests on the belief that these conditions apply in markets or apply at least as much in markets as in monopolies, or do not apply at least as much as under a monopoly, it is open to question whether markets offer any advantage over monopolies in the electricity sector.

Supply and Demand Analysis

In the diagram below, demand can take one of four (typical levels), labelled by reference to the plant type required to operate: baseload (Db); mid-merit (Dm); peaking (Db); or exceeding capacity (Dl). In each case, the market price is given by the variable cost of the marginal plant (Vb, Vm, Vp) or VOLL in the case where load is being shed. The incentive to invest in generation capacity depends on the frequency with which these (typical) demand conditions arise and, in particular, the frequency with which demand reaches Dl, such that load is lost, the price rises to VOLL and generation capacity receives the additional revenue needed to cover fixed costs, as marked by the double-headed arrow.