Electricity and Markets

Richard Green*

University of Hull Business School
Hull HU6 7RX
UK

r.j.green@hull.ac.uk

October 2004

Revised: 09 March 2005

Abstract
Over the last fifteen years, an increasing number of electricity industries have replaced vertical integration with markets as the main method of organising production. Electrical energy is traded in many European and US markets, while the US also has markets for generating capacity. US generators can reduce the cost of complying with environmental regulations by trading emissions of sulphur dioxide, while Europe has just started a carbon dioxide emissions trading scheme. This article discusses the way in which these markets put economic principles into practice. In particular, it shows that several different market designs can provide theoretically equivalent incentives for generators to build capacity, and that emissions trading may have unexpected impacts upon electricity prices.

Keywords: Electricity Markets, Paying for Capacity, Emissions Permits, Tradable Green Certificates
JEL: L94

I. INTRODUCTION

Twenty-five years ago, the electricity industry was largely made up of vertically integrated monopolies. Smaller utilities without their own generation bought their power under contract from a larger firm, or simply paid the tariff that the generator set each year for its power sales. When the utilities with generation wished to trade power among themselves, they typically did so on a split-savings basis. Each utility would report its marginal cost, and the price would be the average of the two figures,

* Support from the Leverhulme Trust, through the award of a Philip Leverhulme Prize, is gratefully acknowledged. I would like to thank the Department of Applied Economics, University of Cambridge, for its hospitality, and Dieter Helm, Janusz Bialek, an anonymous referee and participants at the *Oxford Review of Economic Policy* seminar for helpful comments – the usual disclaimer applies.

thus giving each utility half of the gains from trade. Even where there were "power pools" involving a large number of generators, they operated as clubs, rather than as markets.

Today, many of these monopolies have been broken up, and the electricity industry in a large number of countries has been reorganised around markets. The first wholesale market was established in Chile in 1978, but the great wave of deregulation did not get underway until the 1990s. England and Wales established a spot market, the Pool, in 1990, and Norway liberalised its own market in 1991. In 1996, this became the world's first international electricity market, when Sweden joined the renamed Nord Pool. By then, Australia and New Zealand had created electricity markets, and many others followed them. In the United States, the Pennsylvania-New Jersey-Maryland Interconnector (PJM), which had acted as a club pool since 1927, converted itself into a spot market in 1998, followed by California, New England, New York and Texas.

While electricity spot markets attract a lot of attention, with their highly visible and frequently changing prices, any company that relies on them is taking a large risk. Most companies do most of their trading through longer-term contracts, although this now means contracts lasting not for decades but for years (or even for a single year). Depending on the design of the local spot market, these may be contracts for physical delivery of power, or they may be financial contracts for differences which hedge the spot market price, and reduce the risk of trading in that market.

Over the past twenty years, we have also become more conscious of the pollution caused by electricity generation, and of the challenge of climate change. There has also been a shift from command-and-control regulations for pollution, towards market-based incentives. The United States established a market for sulphur dioxide emissions, a major cause of acid rain, through Title IV of the 1990 Clean Air Act Amendments. This has been followed by a number of regional markets for nitrogen oxides. The European Union still largely relies on command and control to deal with these emissions, but is creating a market for carbon dioxide as part of its efforts to comply with the Kyoto Protocol's targets for the emissions contributing to global warming.

Even before the carbon trading scheme was agreed, most European countries were trying to reduce emissions by promoting renewable generation – wind power, hydro power and the like. At first, these generators received direct subsidies, or sold their power at higher prices to utilities that were required to buy a set proportion of their needs from renewable sources. Schemes like these have the disadvantage that it can be difficult to set the special tariff at the appropriate level to encourage the desired amount of new construction. Some countries have adopted a different system, based on tradable green certificates. Each renewable generator obtains green certificates for its output, and each retailer must hold certificates equal to a specified proportion of its sales. We do not consider these markets further, but they can provide an important stimulus to renewable generators, over and above the impact of carbon trading.

How well are these various markets working? We shall see that a wide range of different market designs have been applied around the world. Does the choice of market design matter, in terms of the prices and investment incentives that it will produce? Can we identify unambiguously good and bad designs, or do some designs

perform well against some objectives, and badly against others? For example, does a price cap that guards against the abuse of market power when there is spare capacity make it hard to remunerate peaking plant and incentivise investment once more capacity is needed? In that case, does the choice of design depend upon the industry's current state? We should remember, however, that making significant changes to market rules should not be done lightly – the implementation costs of writing new contracts and software can be high, and each change creates uncertainty about future changes that can deter investment. It is probably better to choose, and stick with, a market design that will perform well in most circumstances, than to adopt a less robust design that happens to be better suited to the industry's current situation.

The physics of electricity provide some important constraints on the design of an appropriate market, and so section II gives a brief account of these. The next section shows how a spot market with constantly changing prices can allow generators to recover the full costs of all types of generating capacity in equilibrium. Sections IV to VI discuss electricity wholesale markets from around the world, concentrating on the price of energy, different ways of paying for capacity, and transmission pricing. Section VII moves on to discuss long-term contract markets for electricity. Section VIII covers markets for emissions permits, and the article finishes with brief conclusions.

Before we start the main body of the article, it is worth defining a few terms. Generators are at the upstream end of the industry. At the downstream end, retailers sell to consumers. I will use the term "retailers" in this article in order to clearly distinguish between the demand and the supply side of the wholesale markets, although retailers in the UK energy markets (and some others) are known as "suppliers". The wholesale markets are where generators, retailers, and independent traders meet, together with a few large customers who purchase on their own account (a distinction which is not important for our purposes). In the new energy paradigm, both wholesale and retail markets can be opened to competition. Delivering electricity, however, depends upon networks which are natural monopolies, and are usually regulated as such. We distinguish between transmission and distribution networks. Most customers are connected to distribution networks where the voltage is low. The distribution networks are connected to a transmission system that works at much higher voltage. The transmission systems have to be actively managed to ensure that electricity can reach the distribution systems, whereas distribution system management can be more passive – most of the flows just go down the network to consumers.

II. SOME PHYSICS

Electricity is the extreme example of a non-storable commodity. At every moment in time, generation must balance demand, with only a small tolerance for changes in frequency to absorb fluctuations. Generation and transmission equipment alike will be damaged by excessive flows, and is protected by circuit breakers that will cut off the current if they sense a power surge. A single failure can thus cause millions of people to lose power, as large parts of the grid automatically shut themselves down – the North-Eastern United States and Ontario, parts of Sweden and Denmark, and Italy

all suffered from large-scale power failures during the summer of 2003. To minimise the risk of such problems, and to contain them if possible, the electricity system must be tightly controlled. Within each area, there must be a control centre with authority to over-ride individual companies' commercial decisions, if this is necessary to protect the grid. The control centre will have to ensure that there is sufficient spare capacity to cope with the loss of any single element. Transmission investments are planned so that the system can cope with the loss of any two elements. These "N-1" and "N-2" rules imply that there must always be enough spinning reserve – generators that are running part-loaded – to replace the sudden loss of the largest single unit on the system. Similarly, the transmission lines must have enough spare capacity that if one circuit fails and the current flowing down it is redistributed over the rest of the system, none of the remaining lines is overloaded.

Electricity flows cannot generally be steered through the grid, and will follow every feasible route between generators and loads. This means that it is not sufficient to have spare capacity on some route between a generator and a load, in case a circuit is lost – there must be enough spare capacity on every line to cope with the uncontrollable change in flows that it might experience. Sometimes, this means that the system controller will not be able to accept all the flows that market participants would like to schedule, and so part of the transmission system will be constrained. Some generators with relatively low costs, located on the wrong side of the constraint, will have to be replaced by other generators in a better location, but with relatively high costs. In some cases, where the constraint surrounds a relatively small "load pocket", the system operator will have no real choice about which generators to constrain on in this way. In other cases, the system operator has more well-placed plants to choose from. The system operator also has some discretion over whether a constraint exists or not, since the issue is not whether it is physically possible for any more power to flow down a line in the system's current state, but whether there is a significant risk that the system could reach a state in which the additional flow would cause a problem.

III. SOME ECONOMICS

To explain how electricity prices vary over time, we must consider the costs involved. Start with the top right panel (1) of figure 1. The two straight lines show the total cost (per MW) of operating two types of power station for different numbers of hours over a year. The vertical intercept represents the station's fixed costs, while the slope gives the variable cost per MWh generated. The peaking plant (which might represent an open cycle gas turbine, or an old plant nearing retirement) has lower fixed costs but a high variable cost. The base load plant (combined cycle gas turbines are the investment of choice in most countries where gas is available) has higher fixed costs but a lower variable cost. If the plant is needed for more than T hours a year, it is cheaper to build a base load plant, whereas the peaking plant is cheaper for less intensive use. The thicker line segments show the lower envelope of the cost function, for this efficient plant mix.

How much capacity of each type is needed? Move down to panel (2), which shows the load-duration curve. The hours of the year are ranked in decreasing order of

demand, and there are T hours in which the demand is at a level of B GW or above. If the industry has B GW of base load plant, they will all be used for at least T hours a year. If there were any more base load plants, they would be used for less than T hours, and it would have been cheaper to use peaking plants instead.

What about the total capacity? At this point, we need to remember that the demand for electricity, and hence the load-duration curve, depends upon its price. The load-duration curve in panel (2) is drawn on the assumption that the price of electricity is the variable cost of the marginal plant in operation. At a price equal to the variable cost of a peaking plant, the maximum demand for electricity would be D. Peaking plants would not be able to recover their fixed costs, however, if the price of electricity never rose above their variable costs.

In an ideal electricity market, there would be a smooth demand curve for power at each point in time – some consumers have to pay the market price and would be willing to reduce their demand as the price rose. The highest of these demand curves is shown in panel (4) – panel (3) is a reflector which allows us to switch from having capacity on the vertical axis in panel (2), to having it on the horizontal axis in panel (4). The heavy line in panel (4) is the industry's marginal cost curve. There are B GW of base load plant with a variable cost of C_B, and $(K - B)$ GW of peaking plant with a variable cost of C_P. The marginal cost curve becomes vertical at K, the industry's total capacity. If the price were equal to C_P, the maximum demand would be equal to D GW, but since this much cannot be generated, the price must rise to P_R in order to ration demand to capacity.

This leads us to the price-duration curve shown in the final panel, (5). The maximum price is P_R, and the price remains above C_P for the first P hours of the year. At hour P, panel (2) shows that the demand at a price of C_P is just equal to K, there is no need to ration demand to capacity, and the price is equal to the variable cost of the peaking plants. After T hours, demand is low enough to be met by the base load plants alone, and the price drops to C_B.[1] The shaded area just under the top of the load-duration curve (2) represents the electricity that would be demanded at a price equal to variable cost but is not generated, flattening the peak demand to the level of available capacity, K. Note that in a system where rationing by price was not possible, but the available capacity was still equal to K, the shaded area would represent non-price rationing, in the form of "unserved energy" (i.e. power cuts).

Finally, we can return to the top panel (1) to consider the question of cost recovery. The curved line in the bottom left of the panel shows the rate at which plants earn revenue in the highest-demand hours of the year – the line has a steep slope because prices are high. By hour P, this total revenue line has met the thick line giving the total costs of a peaking plant. Since all the peaking plants are in use for at least the first P hours of the year, this allows all of them to cover their fixed costs. Since the marginal revenue for each additional hour's generation up to T is equal to the variable costs of the peaking plants, the total revenue line is superimposed on the total cost line for those plants. After T, the price drops to the variable cost of the base load plants,

[1] The step change in the price-duration curve implies that there is no demand curve passing through the vertical section of the marginal cost curve at B. We could adapt the analysis if some demand curves did pass through this section of the cost curve. In practice, a system with a large number of plants has a nearly smooth marginal cost curve, and no noticeable discontinuities in the price-duration curve.

and so the slope of the total revenue line falls. By the definition of T, however, this is the point where the two types of plants have equal total costs, and so the base load plants will also have total costs equal to their total revenues.

This analysis has been simplified (and more detail is available in Stoft (2002)), but it shows us that if the industry has the right amount and mix of capacity, all plant types can recover their costs from market-based pricing. The key requirement is that peak prices must rise above the variable costs of peaking plant. If we have too little plant, prices will be above this level for longer, and so all types of power station will be paid more than their costs, signalling that entry will be profitable. If there is too much capacity, prices will only exceed variable costs by a small amount, and stations will lose money, encouraging exit. Similar conclusions can be drawn about the mix of capacity – if there are too many peaking plants relative to base load capacity, prices will be high for a greater proportion of the year. This raises the earnings of base load stations above their costs, sending a signal that more are needed.

We should also discuss the pricing of electricity transmission. Ignoring losses for a moment, if there were no transmission constraints, the marginal cost of power could be the same at every point in the system. If there is a binding limit on the flow along some line, however, some more expensive output from the importing side of the constraint must replace some cheaper output from the exporting side. If the price in each region is equal to the marginal cost in that region, then the difference between the two prices represents the marginal cost of the transmission constraint, and the economically correct charge for sending electricity between the two regions. In a meshed network, however, in which there are many possible routes between any two points, it is no longer correct to think of just two prices. If a node on the network is "close" to the constraint – the resistance on the lines between that node and the constraint is relatively low – then Kirchhoff's laws imply that a relatively high proportion of any change in generation at that node would flow across the constrained line. Changes in generation at other nodes will have less impact on the constrained line. This implies that the marginal cost at a node could be thought of as depending on the marginal cost (per MWh) of the constraint, multiplied by the proportion of a 1 MW increase in demand at that node that would actually flow over the constraint.[2]

Transmission losses also lead the marginal cost of power to differ at every point on the system. As electricity flows through the network, an amount proportional to the current squared is lost in heating the wires, and there are also losses in transformers and other pieces of equipment. Because the heating losses depend upon the current squared, the marginal loss is twice the average loss. This implies that the marginal cost of meeting a demand at the importing end of a heavily loaded line can be significantly greater than at the exporting end. At the winter peak, it could be necessary to generate 106 MW in the north of England in order to meet an extra 94 MW of demand in the south-west, if no closer plants were available (NGC, 2004, table 7.4).

If the system has been dispatched in a way that minimises overall costs, then the marginal cost at any node where a generator is running with spare capacity is the

[2] Note that if the node is on the exporting side of the constraint, an increase in demand will *decrease* the flow across the constraint, this proportion will be negative, and the impact of the constraint is to reduce the marginal cost at this node. It is even possible for some marginal costs to be negative!

marginal cost of that generator. However, we would obtain the same number if we calculated the marginal cost of generation at some arbitrary point on the network (known as the "swing bus"), plus the cost in terms of losses of getting the power from the swing bus to our location, plus the impact of additional demand at our location on constraints.[3] The same formula is used to calculate the marginal cost of meeting demand at all other nodes on the system.

IV. ELECTRICITY MARKETS

There are three key questions in the design of an electricity market. First, and probably least controversial, is how to set the price of energy. Second, should there be any specific extra payments for capacity? Third, to what extent (and how) should the prices include transmission effects? This section discusses a range of answers to the first question; the others are answered in the following sections.

Why do there seem to be so many answers to these questions, and such a diversity of market designs around the world? In part, this is due to the complexity of the issues involved – we genuinely do not know the best way of dealing with some of the problems. Sometimes, the equivalence between different market rules comes about because one system will be based on an expectation (particularly of the value of capacity, for example), while another will be based upon an out-turn value. Some markets choose to reward capacity based on prices (letting the market decide quantities) while others use rules based on quantities (so that the market sets the price). While the various rules can be equivalent under theoretical conditions, we do not know how they will perform in practice without experience.

When it comes to practical market design, there can be a tendency to react to perceived flaws experienced in other countries' markets, by moving away from their designs. Policy makers also change their views on which problems are the most important – the UK's first market design used a lot of the industry's former operating procedures in order to minimise the risk of disruption. By the time that the market was re-designed, these rules were seen as encouraging gaming to raise prices, while security of supply was not regarded as such a constraint. After a few years of low prices, low investment, and declining capacity margins, however, security of supply has moved back up the political agenda – will this lead to further market changes?

Moving back to the specific question of how the price of energy is set, we should first ask which transactions it will cover. In a gross pool, all but the smallest generators are required to sell all of their output to the pool, and receive the pool's price for it. In a net pool, generators can agree bilateral trades with retailers. They have to inform the system operator, which will take them into account when drawing up its operating schedule, but they do not pass through the pool from a financial point of view.

In practice, the difference can be more apparent than real. With a gross pool, generators and retailers can sign financial contracts for differences which hedge the

[3] This result holds, in part, because the cost of the constraint is derived from the difference in the marginal cost of generation on each side of it.

pool price, and then bid in a way that ensures their plant output matches the quantity covered by the contracts. As long as they maintain this output level, they face no financial exposure to pool prices. In a net pool, bilateral trades should pay transmission charges in the same way as trades made through the pool.[4]

The world's first electricity market, in Chile, was a net pool in that generators could also sign long-term contracts, and the spot market was only used for the differences between actual and contracted deliveries. However, the economic dispatch that determined actual outputs was based on generators' bids, and the energy price was set equal to the bid of the marginal generator. The bids were not freely made, however, but had to equal the generators' (auditable) costs. Furthermore, there was an adjustment system that came into play if the price in a long-term contract differed from the average spot price by more than 10%.

The second electricity market, in England and Wales, allowed generators to choose their bids freely. Each generating unit submitted a complicated bid with up to five prices and various technical parameters, and the transmission system operator used these to calculate the least-cost schedule that could meet its demand forecast. The program that it used was actually the same program that had been used when the industry was a vertically integrated public corporation, and prices had just been substituted for internal cost estimates. The energy price in each half-hour, the System Marginal Price, was based on the average cost of the marginal unit, or rather the most expensive unit in normal operation. The rules were complex and sometimes produced anomalous prices – spikes caused by the scheduling of a few high-priced MWh became common towards the end of the Pool's life (Offer, 1999).[5]

The world's third electricity market, which started in Norway and has evolved into Nord Pool, covering four countries, had a very different design. Energy prices were set by the intersection of simple supply and demand curves, in which companies just bid the price they would like to pay or be paid for a given quantity. The market is a net pool, and many companies generate their own power, or trade bilaterally. The system operators need to know about these transactions, but they need not enter the market. Elspot is the main market, which sets prices and volumes for an entire day (although separate bids can be submitted for each hour), and closes at noon on the previous day. There are separate short-term markets (Elbas, for example, covers Sweden, Finland and Eastern Denmark) allowing participants to trade in individual hours up to one hour before delivery. After that time, the national system operators take over the task of balancing the markets, using bids for "up-" and "down-regulation". In 2003, 118 TWh was traded on Elspot, compared to consumption of 380 TWh in the four countries. The short-term market Elbas saw much less trade, just 0.6 TWh, compared to consumption of 230 TWh in the area it covered. Because Nord Pool is dominated by hydro-electric generators, the level of rainfall is a key

[4] Some of the support for "trading outside the Pool" in the UK came from companies that believed it would allow them to avoid paying Uplift charges (for keeping the grid stable) in the same way that self-generators (quite reasonably) did.

[5] A price of £865/MWh was once obtained because a large unit submitted a price for starting up of around £12,000 (which was reasonable for a unit of that size) alongside an inconsistent set of technical parameters which caused the computer to attempt to run it at 5% of its capacity for less than two hours. Dividing the start-up price by less than 20 MWh of output produced a very high price. The rules were later changed to allow such anomalous prices to be investigated and sometimes over-ridden.

determinant of prices. Abnormally low rainfall in the autumn of 2002, for example, sent prices to record levels in the winter of 2002/3. This provided incentives for consumers to reduce consumption, and the winter was passed without major incidents, despite low reservoir levels (von der Fehr *et al*, forthcoming).

Australia established a market in Victoria in 1994, which was re-named the National Electricity Market as it expanded to cover the densely populated coastal strip in the south and east of the country. This again used a relatively simple bidding system, and set prices on the intersection of supply and demand curves. One feature was that generators could set negative prices – it costs a lot to turn off and then start up a large coal-fired unit, and so generators sometimes paid to be allowed to run overnight, when demand was low and not all the units were required. The marginal price is calculated every five minutes, and six prices are then averaged to get the half-hourly price used for trading (NEMMCO, 2004).

In the United States, the Pennsylvania-New Jersey-Maryland interconnector and California both opened markets at about the same time, in 1998. California's well-publicised disaster has been covered in many other papers (e.g. Joskow, 2001) and is not discussed here. PJM has been much more successful. Like Nord Pool, it is a net pool, in that generators and retailers can trade bilaterally, although they must inform PJM's independent system operator of their transactions. PJM has a day-ahead market, in which both generators and retailers can submit price-sensitive bids, and in which market-clearing prices and quantities are calculated, taking operational constraints into account. These transactions are financially firm, in that the resulting payments must be made, whatever happens on the day. In the real-time market, PJM schedules plant to meet the actual level of demand, using the cheapest resources as far as possible, and sets prices on the basis of the marginal cost of generation. These prices are used for the differences between previously scheduled transactions (either bilateral or day-ahead) and what market participants actually do. Overall, the energy passing through the spot markets is equal to 40% of average demand, with four-fifths of this in the day-ahead market (PJM, 2004). Table 1 shows that average prices are higher in the day-ahead market than in the real-time market, but that they are also less volatile. Effectively, buyers are paying a premium to avoid the greater risks of real-time purchases.

Table 1: Prices in PJM, 2003

	Day-Ahead	Real-Time	Difference	Difference as % of Real-Time
Mean LMP	$38.72/MWh	$38.27/MWh	-$0.45/MWh	-1.2%
Median LMP	$35.21/MWh	$30.79/MWh	-$4.43/MWh	-14.4%
Standard Deviation	$20.84/MWh	$24.71/MWh	$3.87/MWh	15.7%

Source: PJM (2004) Table 2-27

Most electricity markets are therefore based upon the idea of marginal pricing – the most expensive generator in use sets the price for everyone who has not made a bilateral trade or signed a hedging contract. Great Britain, however, has moved away

from this with the New Electricity Trading Arrangements (NETA) adopted in 2001.[6] To many observers, the Pool had become discredited, and some believed that allowing the marginal generator to set prices for all worsened the market's well-known problem of market power (Green and Newbery, 1997; Green, 2001). The compulsion inherent in a gross pool was also believed to be undesirable, and so NETA is based upon bilateral trading. At Gate Closure, now set one hour before real time, bilateral trading has to stop and companies report their contracts to Elexon, the company responsible for the Balancing and Settlement Code. The National Grid Company (NGC) is responsible for balancing the system in real time, buying and selling power to keep demand in line with supply and resolve transmission constraints. Some of these trades are made at short notice; others can be made in advance, or called in via option contracts signed in advance. NGC records the cost of all its purchases, and the revenues it receives for all of its sales.

Originally, the average cost of NGC's purchases was used to set the System Buy Price. Any company with a negative imbalance – one that had sold more than it generated, or bought less than it consumed – had to pay the System Buy Price for its imbalances. The average revenue from NGC's sales set the System Sell Price. Companies with positive imbalances – those that had generated more than they had sold, for example – were paid the System Sell Price. The System Buy Price was expected to be greater than the System Sell Price, and this was intended to give companies an incentive to balance their positions before Gate Closure, and to penalise deviations from their contracted positions.

In practice, the System Buy Price was much further from the prices in the bilateral markets, and much more volatile, than the System Sell Price. Generators could minimise their expected imbalance payments by deliberately selling less than they expected to produce, and holding plant in reserve in case they suffered an outage. An increase in the level of part-loaded plant reduces the system's efficiency (the economies from pooling reserve exceed the gains from stronger incentives to reduce outages), while NGC faced problems in managing a system where every participant wished to have a surplus of power. The rules were amended so that participants with a "neutral" imbalance – one in the opposite direction from that in the market as a whole, which thus reduces the overall imbalance and could be seen as making NGC's life easier – now pay an imbalance price based upon prices in the short-term bilateral markets. The two imbalance prices are now much closer, reducing the incentive to have an expected surplus of power.

How much impact does the use of pay-as-bid rather than marginal pricing in the short-term market have? Green and McDaniel (1999) show that in a perfectly competitive market with completely inelastic demand, the choice of pricing rule will not affect any generator's expected revenues. The most expensive generator with a chance of being called will bid its marginal costs under both pricing rules. Infra-marginal generators will bid their marginal cost under the marginal pricing rule, and will bid above their costs with the pay-as-bid rule. If they run, these generators will normally be paid more than their bid under the marginal pricing rule, but under NETA, they only receive their bid. The change in bidding exactly offsets the change in pricing rule.

[6] NETA was for England and Wales alone, but the market is to be extended to Scotland in April 2005, with the British Electricity Transmission and Trading Arrangements.

Federico and Rahman (2003) argued that a pay as bid rule is more vulnerable to market power when demand is price-responsive, although their model does not quite follow NETA's principles.[7] Bower and Bunn (2000) used an intelligent agent computer simulation to predict that a pay-as-bid rule would produce worse results than marginal pricing. Prices were low in the first years after NETA was introduced, but this could have been the result of divestitures by the major generators, which reduced concentration in the wholesale market to very low levels. Subsequent mergers have increased concentration, and the real test of NETA as an antidote to market power will be whether price-cost margins remain low in this new environment. The absolute level of prices has risen sharply since 2003, in part because fuel prices have been rising.

In the longer term, NETA may well make entry into the industry harder. Most electricity systems have a liquid short-term market, and because participants can always trade out their positions in this market, it is not necessary to be vertically integrated in order to compete effectively. If the short-term market is illiquid, then a physical hedge between generation and retailing becomes much more important, and unintegrated entry much harder. This would then lead to a less competitive industry and higher prices in the long term. It is certainly the case that many large companies in the electricity industry welcomed NETA (at the same time that the regulator argued that it would make the market more competitive) and that mergers in the industry led to a largely vertically integrated structure by the time the market rules changed.

What can we conclude about the choice of payment system for energy? In a transparent market, the choice between a gross pool and a net pool should be irrelevant to the market outcomes, although a net pool may allow more flexibility in the design of contracts. If bilateral trading is not transparent, however, the gross pool will give market participants more information, and may make trading, entry and exit decisions more efficient. The lack of transparency may be one of NETA's greatest drawbacks. Most other market design choices, however, will trigger offsetting responses by market participants, and so it seems that a wide range of energy payment systems are working well in practice.

V. PAYING FOR CAPACITY

Section III showed that the price of electricity has to rise above the marginal running cost in peak hours if generators are to cover the fixed costs of keeping capacity available. This section discusses some of the ways in which electricity markets raise peak prices to pay for capacity. There are three broad alternatives. First, some markets have no explicit mechanism. Second, some market designs include a payment to capacity, which generally declines as the amount of spare capacity rises, in order to reward capacity most when it is needed most. Third, some systems include an explicit market for capacity, in which electricity retailers are required to contract with a specified amount of capacity (including a margin over their expected demand) or pay a penalty. This sets the quantity directly, whereas capacity payment systems

[7] They assume that all buyers pay the highest accepted bid under either pay-as-bid or marginal pricing, whereas NETA makes buyers pay the average of all the accepted bids, which will be lower.

set the price. Either could produce the same equilibrium values, but the capacity payments system is likely to produce a greater range of outcomes in terms of the level of capacity. Given the importance of avoiding capacity shortages, this could be a significant disadvantage.

Some markets have no explicit payment for capacity. Nord Pool, for example, has none because electricity production in the Nordic countries is dominated by hydro-electricity. The average hydro scheme in the Nordic countries has enough storage capacity to run (at full power) for just under half the hours in the year (Nordel, 2004). The average demand for electricity in 2003 was 73% of the peak demand, however. This means that if the system is to store enough energy to meet the overall demand for MWh over the year, it will invariably have enough "effect capacity" to meet the peak demand in MW. The implication is that market prices need to be high enough to remunerate hydro storage capacity over the year as a whole, rather than to remunerate effect capacity at the system peak. The proportion of thermal plant on the Nordic system may rise in future, however, as much of the available hydro resource has been used. Furthermore, the Swedish transmission operator has been concerned by the lack of reserve capacity to cover peak demands (and water shortages) in its own system (which has a more even balance between hydro and thermal capacity). Svenska Kraftnät has signed medium-term contracts for nearly 2,000 MW of capacity (the country's peak demand is about 26,000 MW), and discussions are underway to find a long-term solution.

The UK, however, has a thermal power system, and NETA is an energy-only market with no special provision for paying for capacity. The implication is that energy prices are believed to provide sufficient reward for capacity. Bids in the balancing mechanism are made very close to real time, when it should be possible to predict accurately whether there is going to be a shortage of capacity. If a shortage is predicted, the most expensive generators will be able to make bids well above their variable costs. This will contribute towards their fixed costs. Newbery et al (2004) point out that even when a few generators are receiving these very high prices, the system buy price is likely to be at much lower levels. This certainly provides much lower incentives for demand-side load management that simply reacts to the market price[8] than a marginal price would, and Newbery et al suggest that incentives to provide generation capacity are also muted.

It is true that the infra-marginal generators are very unlikely to be able to raise their bids to the level of the marginal generators. Green and McDaniel (1999) show, however, that with competitive bidding, no risk aversion, and full information (but random demand), infra-marginal generators will always bid slightly above their costs. Their bid will in fact be equal to the expected revenue (per MWh) that they would have got in a market based on marginal pricing. In the forward markets, arbitrage against the balancing mechanism ensures that the price is also equal to the expected price in a market with marginal pricing. In other words, if a market with marginal

[8] Such "passive" load management may not be the most appropriate type under NETA, in any case. A reduction in demand below the level that the retailer anticipated and contracted for will actually increase its costs, through the unattractive imbalance price. NGC holds tenders for short-notice reserve and similar services, and a number of customers provide these in the form of demand reductions when required.

pricing would produce enough revenue for generators, so will NETA. The impact of imperfect competition or risk aversion on this result, however, is an open question.

Australia also has no explicit payment for capacity, but the market rules include a Value of Lost Load, set at $5,000/MWh, and the price will be set at this level if there is ever a shortage of capacity and customers have to be interrupted. At other times, the VOLL acts as a price cap on generators' bids – if the market is getting short of capacity, they may be able to raise prices towards this level. Allowing the market to set prices at this level brings political risks, even if most of the payments would actually be covered by hedging contracts. Despite this, VOLL was raised to $10,000/MWh in April 2002 in response to fears that the lower level did not give sufficient incentives for reliability. A safety net was added to cap payments, however, linked to their cumulative level over a week (ACCC, 2000).

Most other electricity markets do have a specific payment for capacity. The key choice here is whether the rules should aim to set the price for, or the quantity of, capacity. The Pool in England and Wales had a capacity payment which can be seen as related to the Australian VOLL mechanism. Instead of letting the price rise to VOLL when there was an actual shortage, however, the Pool calculated the risk of a power cut at the day-ahead stage, and then paid all available generators this Loss of Load Probability multiplied by the net value of lost load. The latter was the Value of Lost Load set by the government at £2,000/MWh in 1990 and uprated each year by inflation, less the System Marginal Price (or the station's own bid, for stations that were not scheduled to operate). If the loss of load probability is correctly measured (and experience suggests that the Pool's figures were generally too high), the Pool's overall price should equal the expectation of prices set in the Australian way.

If the Value of Lost Load accurately measures the cost of a power cut, then these payment rules give the expected (Pool) or actual (Australian) value of capacity at the margin. In terms of figure 1, panel (2), this is the value of the unserved energy in the shaded area of the load-duration curve. If the marginal unit of capacity can just cover its fixed costs from the payments it receives, then we would have an efficient outcome.

Capacity payments were often criticised. Some critics asked how a payment which changed every half-hour could act as an incentive for investment in long-lived power stations and thus deliver the correct level of capacity. This missed the point. Adjustments in the level of capacity for next year have to be made by bringing forward, or postponing, the retirement of existing stations, since a station already under construction is effectively committed to arrive, and a station not yet under construction will not be ready in time. It was possible to sign a contract that would hedge capacity payments over the coming year, and thus to lock in the revenues for a station considering retirement. The actual level of capacity payments would almost inevitably differ from the predicted level, but the generator did not need to worry about this when deciding whether to keep its station open. Over the longer term, new generators might expect that the closure decisions on older plant would keep expected capacity payments approximately equal to the fixed costs of such plant. They could then take these revenues into account when deciding whether to enter the market.

A more serious criticism was that the capacity payments could be manipulated, withholding stations from the market. This certainly happened on a small scale, as generators could find it profitable to delay returning a unit after an outage if capacity payments were high – the extra volume would be offset by a reduction in the level of the capacity payments. In the Pool's second year, PowerGen even adopted a strategy of declaring plant as unavailable, raising the level of capacity payments, and then re-declaring it as available in time to receive those higher payments (Offer, 1991)! The Pool Rules were soon changed to prevent this, and the regulator started to monitor capacity declarations. Green (2004a) shows that there is little evidence that capacity withholding affected prices to a significant extent over the years.

The final problem with the capacity payment system was that the calculated loss of load probability became increasingly divorced from reality, due to the way in which stations' past reliability data were used in the calculation. Towards the end of the Pool's life, very high capacity payments were being set at times when there was actually plenty of spare capacity, but some of the stations had suffered outages at unfortunate times during the earlier periods when their reliability had been assessed. In a long-run equilibrium, a capacity payment rule that over-compensates a given level of capacity will encourage the over-provision of capacity until the resulting payments are expected to equal the marginal cost of keeping capacity available. If this does not vary much with the level of capacity, there will be little impact on prices. In the short run, however, an over-generous formula for capacity payments is likely to feed through into the level of consumer prices. If the formula errs on the other side, security of supply could be compromised.

Given this, the trend in the United States has been to establish markets in which electricity companies are required to secure an appropriate amount of capacity, setting the reserve margin directly, and accepting the resulting price. In PJM, for example, every retailer must acquire a specified volume of capacity resources, based upon their forecast (annual) peak load, scaled up to take account of reserve requirements and generator reliability. Retailers who do not have sufficient capacity must make a deficiency payment, based on the annual fixed costs of a peaking plant (a daily penalty of $170.96/MW was equivalent to $62/kW-year in 2003/4). This gives peaking capacity an additional revenue stream, and ensures that every retailer is bearing a share of the costs of reserve capacity. The reserve margin incorporated in the capacity requirement, and the deficiency payment (which affects retailers' willingness to pay) can be adjusted to encourage the market to supply the level of capacity which policy-makers believe is desirable. If the rules are adjusted too often, of course, then companies may lose confidence in the market as a reliable source of revenue, and it will become less effective.

The choice between price- and quantity-based ways of paying for capacity can be viewed in the light of Weitzman's (1974) discussion of the choice of policy instrument in the presence of uncertainty. Weitzman showed that where costs and benefits were uncertain, a quantity standard would be preferred when the slope of the marginal benefit function was steeper than the slope of the marginal cost function. The marginal cost of providing capacity is likely to be fairly flat over the relevant ranges, whereas the marginal benefit of extra capacity rises quickly as the level of

spare capacity falls.[9] This suggests that capacity markets should be favoured over price-based systems, at least as long as the evolution of demand is taken into account when the quantity required is determined.

Capacity markets are also likely to have political advantages over capacity payments. Systems based upon prices alone tend to need occasional very high prices if they are to create enough revenue, and these create financial and political risks, even with hedging contracts. Generators may not believe that prices will be allowed to rise to the necessary levels. A generator must face the risk that it will not be available during the price spike, losing the associated revenue, and will be even worse off if it has contract payments to make. The indirectness of the link between the high prices and improved security of supply is also likely to be a disadvantage as investment needs move up the political agenda. A capacity market has a direct link to security of supply, which makes it easier to justify the resulting prices. Signals to investors can be clearer when there is an obvious demand for capacity, even though the time lags involved in building plant can help to create investment cycles.

Markets with no explicit mechanism for ensuring enough capacity may work in theory, or where there are large amounts of price-responsive demand. If an energy-only market with a low price elasticity of demand encounters a season of high demand or low supply, as happened in California, it is likely to be vulnerable to power cuts and politically unacceptable prices. Overall, the costs of insufficient capacity greatly outweigh those of having too much. Many electricity systems need a lot of investment over the next decades, as plant built in the 1960s and 1970s retires, and as renewable generators replace conventional stations. However, the costs and likely profitability of this investment remain uncertain. In this context, the quantity-based signal of a capacity market is more likely to produce an acceptable level of investment than the price-based signal of capacity payments.

VI. TRANSMISSION PRICING

Section II pointed out that constraints and losses on the transmission system mean that the cost of electricity can vary significantly from place to place. The British, however, have generally dealt with geographical issues by ignoring them. The system operator could not ignore transmission constraints, of course, and had to buy more output from generators inside import constraints, while selling some of their output back to generators on the other side of the constraints. In some markets, this is known as counter-trading. The net cost of this (since the generators constrained on were paid more than the system operator received from the generators constrained off) was recovered from all customers. From 1994 onwards, the system operator was given a financial incentive to keep these costs down. Transmission losses were never used to set regional prices – a rule change to introduce this was blocked by generators who stood to lose from it. When the regulator was given more power over the market's rules, under NETA, the rules were changed to take account of transmission losses, but before this could be implemented, it was blocked by the government, which believed

[9] The number of hours for which the marginal unit of capacity can expect to run increases as the level of capacity falls, as can be seen from the shape of the load-duration curve in panel (2) of figure 1.

that it would hamper the expected (and desired) growth of renewable generation, located mainly in the north of the UK.

Nord Pool has two separate approaches for dealing with transmission constraints. When the transmission lines between two countries (or between regions within Norway) are congested, Nord Pool calculates separate prices for each side of the constraint, and cross-border flows are charged a transmission fee equal to the price difference. Congestion within a region is dealt with by counter-trading. This market splitting procedure is well understood within the market and produces reasonable price signals, although Glachant and Pignon (forthcoming) argue that the rules can provide perverse incentives for the system operators. Congestion within a country is dealt with by counter-trading, which forces the system operator to incur some costs, whereas cross-border congestion brings in revenue from the price differentials. The system operators clearly have an incentive to declare that the constrained lines are at the borders, even if the constraint is actually within a country.

PJM uses nodal pricing, and so prices across the network vary whenever a link is constrained, to reflect the cost of the constraint. Companies have to pay the difference in nodal prices between the point where they put energy into the system and the point where they take it out, whether they are trading in the PJM market or bilaterally. An earlier system which allowed bilateral transactions to escape the impact of congested transmission quickly collapsed when everyone switched to bilateral transactions (Hogan, 1998). Because PJM's system can produce large geographical price differences, participants can hedge these with Financial Transmission Rights (FTRs), known as transmission congestion contracts when first proposed (Hogan, 1992). These effectively pay the difference between a nodal price and the price at the market's swing bus, and so a participant holding FTRs equal to their actual generation can ensure that they receive the swing bus price.

Although other market designs have been tried, it looks as if most restructured markets in the United States are evolving towards the PJM model. New York and New England chose similar designs when they first restructured their markets, and California's tortuous re-design process is moving in the same direction. Texas opened a state-wide wholesale market in 2002, based primarily on bilateral trading, and using a zonal model (like Nord Pool) to manage congestion. This state, too, is now redesigning its market to adopt nodal pricing. The PJM design is also the basis of the Standard Market Design proposed by the Federal Energy Regulatory Commission (2002). The politics of states' rights, however, mean that FERC is unable to impose this market on states that do not wish to restructure. Most of the states that had below-average electricity costs have no wish to adopt market-based prices that would probably be higher. Economic efficiency can come a poor second to practical politics.

Transmission pricing is the area where the choice of market rule most clearly makes a difference to the market outcomes. To put it at its most starkly, some customers and generators will gain from cost-reflective transmission pricing, and some will lose. When the losers are aware of this (and generators certainly will be) they are likely to do everything they can to prevent the imposition of these charges. The question then is whether the overall gains are worth the costs of pushing through the changes. Green (2004b) estimates that the gains from adopting locational transmission prices in

England and Wales could equal about 1.5% of generators' revenues in a competitive market, and would be higher in the presence of market power. Whether such a policy is worth pursuing is a normative question.

VII. ELECTRICITY CONTRACT MARKETS

Electricity spot prices are volatile. They can accurately signal the changing marginal cost of power, but most customers do not change their consumption in response to real-time price signals – most are not equipped to receive such signals, or motivated to change their behaviour if they did. Retailers can reduce the impact of spot price variations – on both their customers and themselves – by signing contracts lasting for longer periods. In a market with a gross pool, these have to be financial contracts for differences, while a net pool allows either physical or financial contracts. A combination of a contract and an appropriate set of bids allows a company to fix the cost of a given volume of electricity, while responding at the margin to the spot market price. Because contracts reduce the importance of the spot price for the company's profits, they can act to mitigate market power. Some markets deliberately imposed vesting contracts, with prices and volumes determined before restructuring took effect, in order to curb market power. In California, the reluctance of the major utilities to sign longer-term contracts made them more vulnerable to price increases caused by increasing costs and the exercise of market power, and contributed to the state's disaster.[10]

A successful forward market requires liquidity. In the UK, multi-year contracts related to coal purchases and to the construction of new gas-fired power stations took up much of the market, so that most companies had little need for additional contract trading until the second set of coal contracts expired in 1998. This discouraged outsiders from entering the markets, and liquidity was low. While the contracts were meant to be hedges on the underlying market of the Pool, they had a significant impact on the way that the generators bid their plant there. Although developments in the Pool were often driven by the contract market, it was perceived as complicated and vulnerable to exploitation by the dominant generators. This also made independent traders reluctant to enter the contract market. Nord Pool, in contrast, has a simple price-setting mechanism and a much more competitive market structure. This creates confidence and liquidity, producing a virtuous circle, so that the volume traded in forward and futures markets is now several times the annual electricity output of the member countries.

Liquidity could be a problem in a market with volatile transmission prices. Companies might be concerned to trade for delivery at their own location, since trading elsewhere would not provide an adequate hedge, but this would then fragment the market. PJM's approach to this problem is to create a number of trading hubs at different points on the network, at which liquidity can be concentrated. The hub price

[10] It is fair to write of the utilities' reluctance, because the California Power Exchange did create a market for forward contracts, and the Public Utilities Commission did agree that the utilities could treat the cost of buying a proportion of their power on this market as an allowable expense for regulatory purposes. Southern California Edison made much more use of this market than Pacific Gas and Electric, and it was the latter company that declared bankruptcy (Blumstein *et al* 2002).

is a weighted average of the prices at a number of nearby nodes, with pre-specified weights. Physical market participants can then use financial transmission rights to hedge the difference between the price at the trading hub and at their location.

The other problem facing contract markets is that if generators are to use them to finance investment in new power stations, the contracts will need to last for many years. Retailers can be willing to sign such contracts if they are reasonably sure that they can pass on the costs to their consumers, and this was the case when retailers had monopoly franchises. Once retail markets have been opened to competition, however, retailers face the risk that retail prices will be closely linked to the current wholesale prices. If wholesale prices fall, this will leave companies unable to pass through the cost of their contracts. This would then make them reluctant to sign long-term contracts with generators, enhancing market power in generation both directly and by making entry harder (Newbery, 2002). Furthermore, Green (2004c) shows that a quantitatively significant effect can occur, even if retail markets are relatively uncompetitive, which seems to be the case at present, as argued by Waddams Price in this issue.

VIII. EMISSIONS MARKETS

Many types of electricity generation can cause pollution. The risk of radioactive pollution from nuclear power stations was probably the first to be recognised. Acid rain, caused by emissions of sulphur and nitrogen oxides, came to prominence in the 1980s. In the 1990s, the world became increasingly aware of the problem of global warming and the impact of carbon emissions.[11] Reducing these emissions is now a major policy aim in the EU.

Traditionally, environmental regulation has taken a "command and control" approach. Regulators have decided on an appropriate level, or rate, of emissions, or a technological standard for emissions reduction equipment, and ordered companies to comply with it. The problem with this approach is that the regulators will almost always know less about each company's cost of reducing emissions than the company itself does. If, as is very likely, this cost varies significantly between companies, then a uniform standard will not minimise the cost of achieving a given total level of emissions. Regulators may try to take costs into account when setting their standards (the UK used to require the "best available technology not entailing excessive cost", for example) but the problem of asymmetric information remains.

Emissions trading uses a market to reveal which companies can reduce their emissions at least cost. After unsuccessful experiments in the 1970s, the US Environmental Protection Agency allowed refineries to trade the lead content in petrol between 1983 and 1987, helping them to phase in a reduction in the average content from 1.1 grams to 0.1 gram per leaded gallon. Refiners that were able to reduce their lead content ahead of schedule could sell their spare permits to others, allowing them

[11] Climate change policy is discussed at length by Helm (2003) and the other papers of Vol. 19, No. 3 of this review

more time to reach compliance. The cost of meeting the new standard was reduced, without using more lead on average than the standard implied.[12]

In 1990, Congress passed Title IV of the 1990 Clean Air Act Amendments, which established a market for sulphur dioxide emissions from power stations. The market started in 1995, covering the largest stations (Phase I), and was extended to all stations in 2000 (Phase II). Annual emissions, 16 million tons in 1985 and 14 million tons in 1994, had to fall to under 10 million tons in 2000, and about 9 million tons by 2005. Emissions from the affected stations fell by more than one-third in the programme's first year, and were roughly one-third below the limit over the five years of Phase I. This meant that nearly 12 million tons of sulphur dioxide had been banked by the start of Phase II. This gave generators more time to adjust to the tougher cap in Phase II. Ellerman et al (2000) calculate that trading between plants within Phase I saved $1.8 billion (discounted to 1995), and that trading between plants within Phase II will save $17 billion by 2007. Banking permits between Phase I and Phase II will have saved a further $1.3 billion. Overall, the cost of reducing emissions was cut by 57%.

Plants were allowed to opt in to Phase I, and the rules for allocating permits were clear enough to allow rational decisions to be made. Some plants opted in because their abatement costs were low and so they could profit by reducing emissions ahead of schedule. Their participation will not have raised total emissions over the life of the market. Other plants opted in because their emissions were already below their potential allocation of permits, and so they would have a surplus to sell, even if they did nothing. Allowing these plants to join the market does raise total emissions when they sell their spare permits, but Montero (1999) estimates that the effect was less than two per cent of total predicted emissions between 2000 and 2009. Offsetting this is the clarity produced by having simple rules for allocating the permits. In particular, the allowances given to generators were based upon their past behaviour. This is equivalent to a lump-sum transfer, and does not affect their future behaviour.

The European Union (2003) has adopted an Emissions Trading Scheme (ETS) to help it reduce emissions of carbon dioxide. This covers all plants in some industrial sectors (coke ovens, oil refineries, and the manufacture of iron and steel, pulp and paper, cement, glass and ceramics) and all plants with a thermal input of more than 20 MW. Affected plants must surrender allowances equal to their emissions of carbon dioxide during a calendar year within four months of the year end, or pay a penalty. The first phase runs from 2005 to 2007, the second phase from 2008 to 2012, and subsequent phases will be for five-year periods. As with the Acid Rain program, the first phase has something of the nature of a lead-in, since Member States can exclude some of their plants from the scheme, and the buy-out payment for non-compliance is €40/tonne, rather than €100/tonne in Phase II. Banking is automatic between the second and subsequent phases, whereas the Directive allows Member States to choose

[12] To the extent that a refinery was already ahead of the standard, allowing it to sell its spare permits does increase emissions, compared to business-as-usual for that refinery and no access to permits for would-be-buyers. The earlier schemes had involved bureaucratic mechanisms that attempted to prevent companies from getting spare permits "for nothing", but this complexity had been one reason for their relative failure (Ellerman et al, 2003)

whether to convert unused allowances from the first phase into second-phase allowances.[13]

The allocation of allowances is left to the Member States, subject to Commission approval, and a requirement that at least 95% (90%) are given away free of charge in the first (second) phase. In practice, most Member States have given away all of their allowances, apart from a reserve for the use of new entrants. If entering an affected industry (almost) automatically endows a firm with valuable allowances, while shutting a plant means that it has to give up all future allowances, this could distort decisions on whether to open and shut plants. The scheme should still ensure that operating decisions at the margin take the new cost of emissions into account, however.

How large is that cost? Ratcliffe on Soar, a large coal-fired station, emitted 0.9 tonnes of CO_2 for each MWh that it generated in 2003, while Connah's Quay, a CCGT from the mid-1990s, emitted 0.4 tonnes per MWh (PowerGen, 2003). If CO_2 allocations cost €10/tonne, then this will add about £6/MWh to the marginal cost of the coal-fired plant, and about £2.75 to the marginal cost of a CCGT. The impact on the price of electricity in any half-hour depends upon which plants are marginal, and whether the ETS changes this. (If a low coal price offsets the higher thermal efficiency of CCGT plants, then coal plants might be cheaper without the ETS, but their higher emissions could make them more expensive than gas plants once the ETS starts.) It also depends upon how competitive the market is – oligopolies typically absorb part of any increase in costs. Electricity generation in Great Britain is reasonably competitive at present, however, and recent estimates of cost pass-through include 90% (Carbon Trust, 2004) and 100% (Ilex, 2004). Ilex estimate that the overall wholesale price of electricity will rise by £3.70/MWh during phase I of the scheme.

What will this imply for generators' profits? Keats Martinez and Neuhoff (2004) suggest that gas-fired generators will gain, even if they have to buy all of their allowances. First, they will run more. Second, in the hours for which they are infra-marginal, the price will normally be set by generators with higher emissions levels, and so the ETS will raise prices by more than their costs. When the plants are marginal, the ETS affects the market price by just as much as their marginal costs, and so they do not lose.

Efficient coal-fired generators could win or lose, in principle, if they had to pay for all their allowances. Keats Martinez and Neuhoff assume that these generators would have been infra-marginal in medium-demand hours, in the absence of the ETS, when the price would have been set by gas-fired stations. This price therefore rises by less than the coal-fired stations' costs: either the gas-fired stations are still on the margin in those hours, and the price rises by their marginal emission costs, or the coal-fired stations are now on the margin, losing their infra-marginal rent entirely. Offsetting this, there will be some high-demand hours when generators with even higher emissions are setting the price, and so the ETS raises the efficient firms' margins. A priori, we cannot tell which effect will dominate. Simulations in the paper, however, show that the coal-fired generator's margin over fuel, emissions and operations and

[13] Presumably it will only take one Member State to agree to do so to ensure that all unused allowances are transferred to organisations in that State and duly converted at the end of the first phase.

maintenance costs would more than halve, from £27.5/kW-year to £12.4/kW-year, if it had to buy all of its allowances. If the plant was allocated two-thirds of the allowances it needed for free,[14] however, its margin would rise to £35.3/kW-year. The gas-fired plant's margin would rise from £27.2/kW-year to £32/kW-year with no free allocation, or £48/kW-year with an allocation based on past emissions.

What about the long-run impact, once capacity can adjust fully? Figure 2 shows the analysis for an example in which peaking plants have higher emissions than base load plants. Panel (1) is a repeat of the top panel of Figure 1, showing the total costs of the two types of plants, for different numbers of hours of operation. In panels (2) and (3), the higher marginal costs imposed by carbon permits imply that the slope of the total cost lines will increase. If the permits are auctioned, then the lines will pivot around their previous vertical intercepts. If some permits are grandfathered, then the intercept will shift down by the value of the free permits. The break-even point is where the two lines cross. In panel (2), base load plants are given enough free permits to run for about $2T$ hours (T being the highest number of operating hours for which it would be economic to build a peaking plant). In panel (3), peaking plants are given enough free permits to run for about $T/2$ hours. Assume that the grandfathering system means that any new plant is given the same allocation as an existing plant, and that any plant that closes receives no more allowances – this is the general thrust of most national allocation plans under the ETS.

The thick line in panel (4) shows the envelope of efficient costs if permits are auctioned. In a sustainable competitive market, these must also equal the plants' revenues, as described in section III. The panel therefore also shows the original cost envelope from panel (1) of Figure 1 (the thin line with a kink at T), which should be interpreted as the revenue required without carbon trading. Without carbon trading, it was efficient for peaking plants to run for up to T hours a year. With carbon trading, these carbon-intensive plants should run much less, for T' or less. The price in each hour is given by the slope of the cost/revenue line. Prices are clearly higher for the peak hours to the left of T', and the off-peak hours to the right of T. In the intervening hours, prices might rise or fall, depending on a comparison of the cost of a peaking plant without permits (originally marginal) and a base load plant with permits (marginal now).

In panel (5), the impact of a grandfathered allocation of permits is shown, on the assumption that the base load plants are given more free permits than the peaking plants. (The value of the free permits is equal to the vertical distance between the "pemits auctioned" and "permits grandfathered" lines in panels (2) and (3), and the base load plants clearly get more permits in this example.) The free permits offset so many of the fixed costs of a base load plant that it is only efficient to use a peaking plant if its running time is very short - T″ is close to the origin. The peaking plant's net fixed costs, that have to be recovered from prices above its marginal running cost, are also very low. This means that fewer hours of very high prices are required, although prices are higher than before in the next group of hours, when they are equal to the running costs of the peaking plant. As panel (5) is drawn, however, it should be clear that for plants running for around T hours, the revenues that they need are well

[14] Note that this hypothetical allocation is equal to the plants' average emissions in an earlier period when its output was lower, and does not depend upon its actual emissions in the modelling exercise.

below the revenues needed without carbon trading. In other words, the price averaged over the highest *T* hours of the year will be lower in the presence of carbon trading. Over the year as a whole, however, time-weighted average prices will be higher, since the free allocation to a base load plant is not sufficient to allow it to run for the entire year without having to buy some permits, raising its total costs.

The specific predictions of this example depended upon the peaking plants having a higher carbon intensity, and a smaller free allocation, than the base load plants. Change these assumptions, and the results will change. The aim, however, is to show that the ETS will affect entry and exit decisions, and is quite capable of reducing electricity prices in some periods, even though we can expect generators' overall revenues to rise.

Is the ETS an efficient response to the problem of climate change? Pizer (2002) uses Weitzman's (1974) results on the choice between suggests that quantity- and price-based regulation, and the fact that the marginal damage function of carbon emissions is almost horizontal (although its level is uncertain),[15] to argue that a tax-based scheme would be more efficient. He does point out, however, that a permit scheme allows governments far more flexibility in allocating the costs and rents involved in reducing carbon emissions, and avoid the political baggage of a new tax. He also suggests that the best policy would combine a permit system with a buy-out charge. This would allow agents to exceed the permitted level of emissions if the out-turn cost of reductions proved to be too high, while letting the government distribute rents when it allocated permits. The ETS does indeed include such a buy-out price, although at a level (€27/tonne of carbon) which is well above the tax levels that Pizer considers optimal for the first decades of this century. The ETS is probably best seen as a reasonable response to political constraints. Its full effectiveness may well depend upon the way in which permits for Phase II and afterwards are allocated, which has not yet been decided.

IX. CONCLUSIONS

Electricity companies have been participating in an ever-growing number of markets over the past fifteen years. The motivation for this is that price signals created in a competitive market are believed to give better incentives than vertically integrated companies have had in the past. The analytical results in this article show that markets are capable, in theory, of signalling the marginal cost of power and of providing the right incentives to build capacity. The examples from around the world show that many different market designs have been used in the attempt to translate these results into practice.

In the United States, there seems to be a consensus that an energy market based upon nodal marginal cost prices, coupled with a market for capacity, is the best structure for a liberalised electricity industry. (There is no consensus about whether to liberalise or

[15] The damage from climate change is related to the stock of greenhouse gases in the atmosphere, and the flow of emissions of carbon dioxide only changes this very gradually. I am indebted to an anonymous referee for directing me to this reference.

not, however.) In Europe, there are few capacity markets, and transmission pricing is much less sophisticated. Europe has moved ahead of the United States, however, in implementing markets for emissions of carbon dioxide, and for supporting renewable generators.

How much do these choices matter? We do know that choices about transmission prices can have large impacts on how (and whether) prices vary over space, and can lead to large transfers between agents. There are worthwhile efficiency gains from implementing prices that signal marginal costs, but they are almost always much smaller than the transfers.

In advance of real time, electricity can be traded in a variety of ways, but real-time operation must be delegated to an operator who can keep the system stable. The way in which these real-time operations are paid for will affect all the other electricity trades through arbitrage, since waiting until real time and paying the prices determined then is always an option for demand. Almost all electricity markets set real-time prices upon the basis of the marginal plant in operation, which should send transparent and efficient price signals through the market. In the UK, the designers of NETA believed that agents should be deterred from real-time trading, and that creating a less transparent market without marginal price signals would produce better results. It is hard to believe that this could be the case. The biggest impact of NETA is probably not in short-term trading, but in the trend towards vertical integration between generators and retailers, which is a response to the lack of transparency in the market, and likely to make independent entry much harder. That will affect the long-term structure of the market, and the evolution of capacity.

Theoretical arguments suggest that getting the right level of capacity does not depend upon having a capacity market, and that price-based systems can send the right incentives. The length of an investment cycle, relative to the few years since most countries liberalised their electricity markets, means that we do not have enough experience to be sure that these arguments will work in practice. Many OECD countries started their reforms with a surplus of generation capacity, ensuring that it would be some time before much investment was needed – although some experienced investment booms in any case. With the correct incentives, demand response can act as a substitute for at least some spare capacity, at a rather lower cost. Given the importance of electricity to the modern economy, however, we need to ensure that the system operators can expect to have enough capacity available to them. A well-designed capacity market is likely to produce more reliable results than using energy prices alone.

BIBLIOGRAPHY

ACCC (2000) *Determination: Applications for Authorisation: VoLL, Capacity Mechanisms and Price Floor,* File no. C1999/865, Canberra, Australian Competition and Consumer Commission

Blumstein, C., L.Friedman and R.J.Green (2002) "The History of Electricity Restructuring in California" *Journal of Industry Competition and Trade*, Vol 2, Nos 1-2, pp 9-38

Bower, J. and D.W.Bunn (2000) "Model-based comparisons of Pool and Bilateral Markets for Electricity" *The Energy Journal*, vol. 21 no. 3 pp 1-29

Carbon Trust (2004) *The European Emissions Trading Scheme: Implications for Industrial Competitiveness,* London, The Carbon Trust

Ellerman, A.D., R. Schmalensee, P.L. Joskow, J.P. Montero, and E. Bailey (2000) *Markets for Clean Air: The U.S. Acid Rain Program* Cambridge, Cambridge University Press

Ellerman, A.D., P.L. Joskow and D. Harrison (2003) *Emissions Trading in the U.S. Experience, Lessons and Considerations for Greenhouse Gases*, Arlington, Pew Center on Global Climate Change

European Union (2003) *Directive 2003/87/EC of the European Parliament and of the Council of 13 October 2003 establishing a scheme for greenhouse gas emission allowance trading within the Community and amending Council Directive 96/61/EC*, Brussels, Commission of the European Communities

Federal Energy Regulatory Commission (2002) *Notice of Proposed Rulemaking, Docket No. RM01-12-000, issued July 31, 2002* (the "Standard Market Design"), Washington DC, Federal Energy Regulatory Commission

Federico, G. and D. Rahman (2003) "Bidding in an Electricity Pay-as-Bid Auction" *Journal of Regulatory Economics*, vol. 24, no. 2, pp. 175-211

von der Fehr, N-H M., E.S. Amundsen and L. Bergman (forthcoming) "The Nordic Market: Signs of Stress?" *The Energy Journal*

Glachant, J.-M. and V.Pignon, (forthcoming) "Nordic Congestion's Arrangement as a Model for Europe? Physical constraints and Economic Incentives", *Utilities Policy*

Green, R.J. (2001) "Markets for Electricity in Europe", *Oxford Review of Economic Policy,* Vol 17 no 3, pp 329-345

Green, R.J. (2004a) *Did English Generators Play Cournot? Capacity Withholding in the Electricity Pool*, CMI Working Paper EP41, University of Cambridge

Green, R.J. (2004b) *Electricity Transmission Pricing: How much does it cost to get it wrong?* CMI Working Paper EP63, University of Cambridge

Green, R.J. (2004c) *Retail Competition and Electricity Contracts*, CMI Working Paper EP33, University of Cambridge

Green, R.J. and T.M.McDaniel (1999) *Expected Revenues in the Balancing Market: Equivalence between pay-as-bid and SMP* DAE working paper 0002, University of Cambridge

Green, R.J. and D.M.Newbery (1997) "Competition in the Electricity Industry in England and Wales" *Oxford Review of Economic Policy* vol. 13 no. 1, pp 27-46

Helm, D. (2003) "The Assessment: Climate change policy" *Oxford Review of Economic Policy* vol. 19 no. 3, pp 349-61

Hogan, W.W. (1992) "Contract Networks for Electric Power Transmission", *Journal of Regulatory Economics,* vol. 4, no. 2, September, pp 211-242

Hogan, W.W. (1998) *Independent System Operator: Pricing and Flexibility in a Competitive Electricity Market*, mimeo, Harvard University

Ilex (2004) *Impact of the EU ETS on Electricity Prices: A report to DTI*, Oxford, Ilex Energy Consulting

Joskow, P.L. (2001) "California's Electricity Crisis", *Oxford Review of Economic Policy*, vol. 17 no. 3, pp 365-388

Keats Martinez, K. and K.Neuhoff (2004) *Allocation of Carbon Emission Certificates in the Power Sector: How generators profit from grandfathered rights*, CMI working paper EP49, Department of Applied Economics, University of Cambridge

Montero (1999) "Voluntary Compliance with Market-Based Environmental Policy: Evidence from the U.S. Acid Rain Program", *Journal of Political Economy*, vol. 107, no. 5, pp 998-1033

NEMMCO (2004) *An Introduction to Australia's National Electricity Market*, Melbourne, National Electricity Market Management Company

National Grid Company (2004) *Seven Year Statement March 2004*, Warwick, National Grid Company

Newbery, D.M. (2002) "Problems of liberalising the electricity industry" *European Economic Review*, vol. 46, pp 919-927

Newbery, D.M.G, W. Nuttall and F. Roques (2004) *Generation adequacy and investment incentives in Britain: from the pool to NETA,* CMI working paper EP58, Department of Applied Economics, University of Cambridge

Nordel (2004) *Annual report 2003,* (available from http://www.nordel.org) Vällingby, Nordel

Offer (1991) *Pool Price Enquiry* Birmingham, Office of Electricity Regulation

Offer (1999) *Pool Price: A Consultation by Offer, February 1999* Birmingham, Office of Electricity Regulation

Pizer, W.A. (2002) "Combining price and quantity controls to mitigate global climate change" *Journal of Public Economics*, vol. 85, no.3, pp. 409-34

PJM (2004) *State of the Market 2003*, Norristown, Pennsylvania, PJM

Stoft, S.E. (2002) *Power System Economics: Designing Markets for Electricity* Chichester, Wiley

Weitzman, M.L. (1974) "Prices versus quantities" *Review of Economic Studies,* vol. 41 no.4, pp 477-91

Figure 1 - The determination of electricity prices

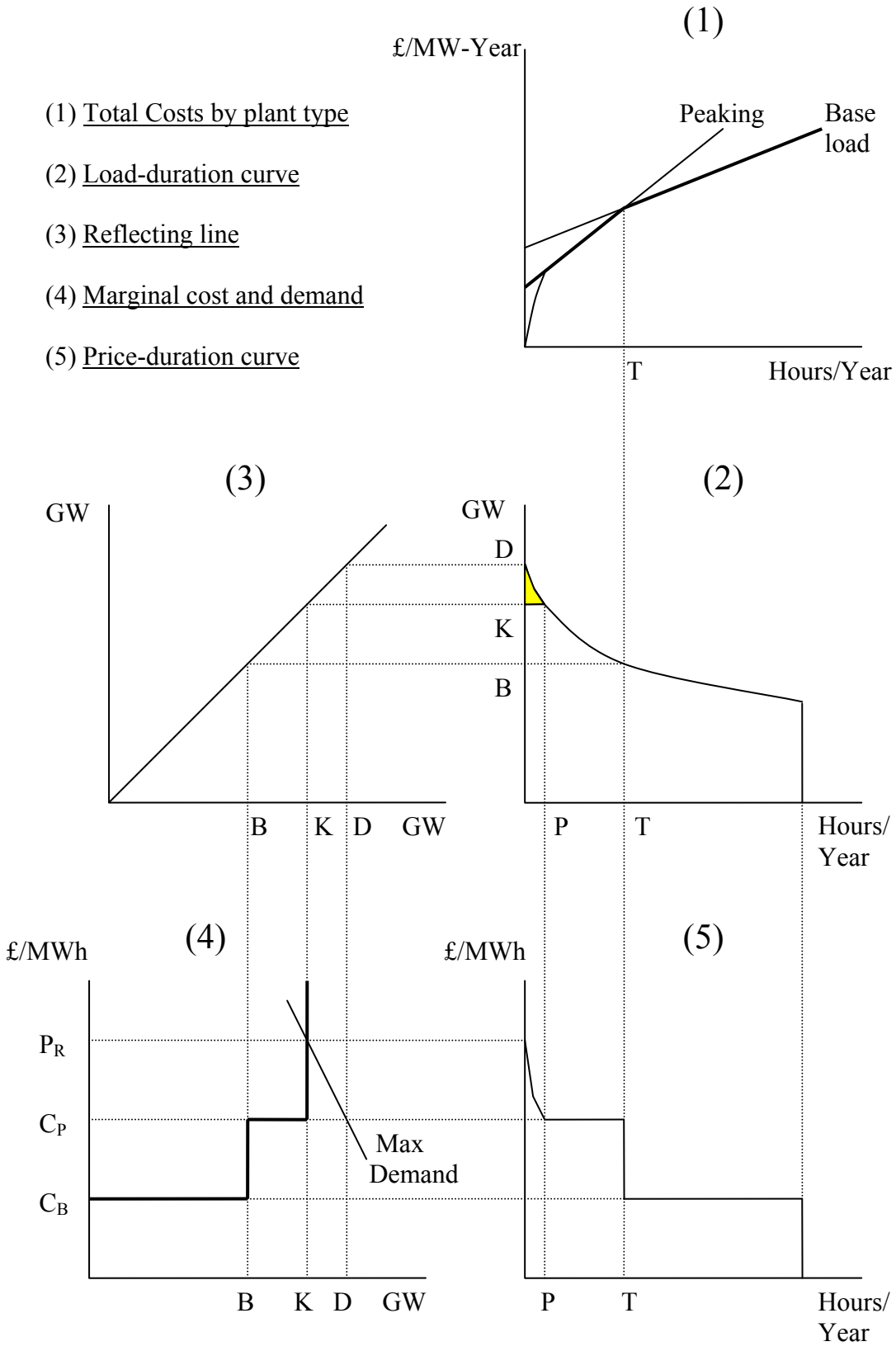


Figure 2 - The impact of a carbon trading system

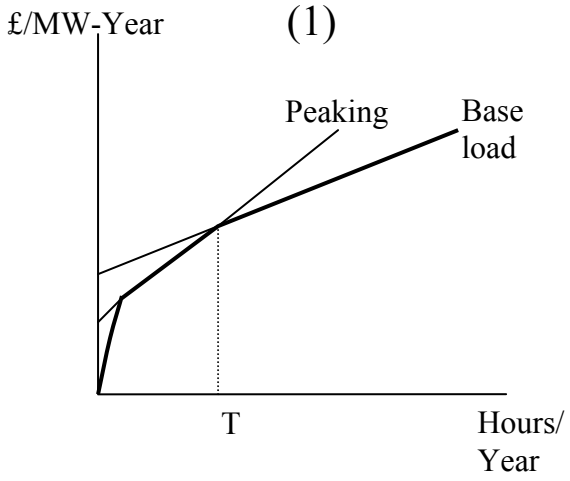
(1) Costs before carbon trading

(2) Total Costs of base load plant

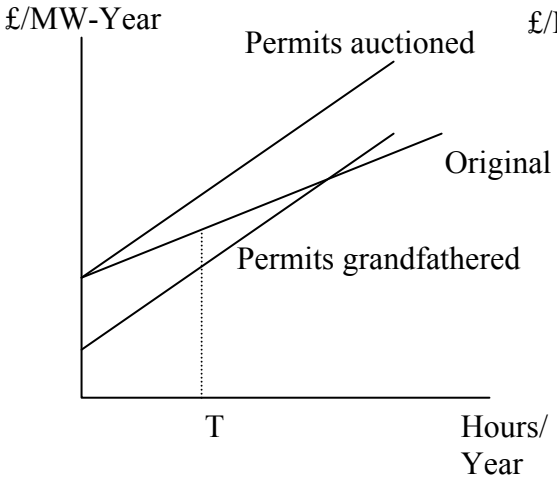
(3) Total Costs of peaking plant

(4) Costs with auctioned permits

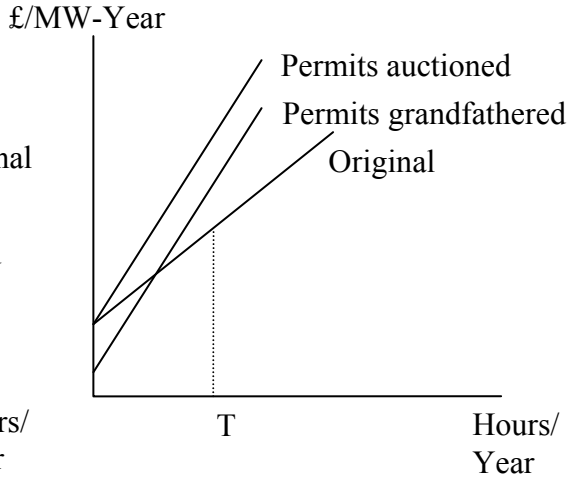
(5) Costs with grandfathering



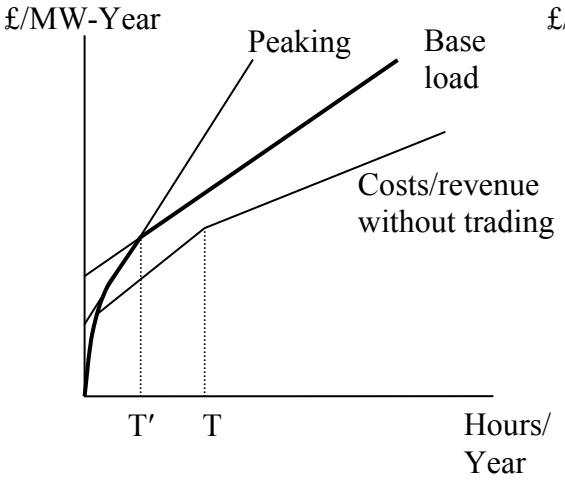
(2)



(3)



(4)



(5)